Food for Thought …
Uncertainty of Testing Methods – What Do We (Want to) Know?

Martin Paparella[1], *Mardas Daneshian*[2], *Romana Hornek-Gausterer*[1], *Maximilian Kinzl*[1], *Ilse Mauritz*[1], *and Simone Mühlegger*[1]

[1]Environmental Agency Austria, Vienna, Austria; [2]Center for Alternatives to Animal Testing-Europe, University of Konstanz, Konstanz, Germany

Summary

It is important to stimulate innovation for regulatory testing methods. Scrutinizing the knowledge of (un)certainty of data from actual standard in vivo methods could foster the interest in new testing approaches. Since standard in vivo data often are used as reference data for model development, improved uncertainty accountability also would support the validation of new in vitro and in silico methods, as well as the definition of acceptance criteria for the new methods. Hazard and risk estimates, transparent for their uncertainty, could further support the 3Rs, since they may help focus additional information requirements on aspects of highest uncertainty.

Here we provide an overview on the various types of uncertainties in quantitative and qualitative terms and suggest improving this knowledge base. We also reference principle concepts on how to use uncertainty information for improved hazard characterization and development of new testing methods.

Keywords: hazard uncertainty, assessment factors, toxicological testing methods, 3Rs

1 Introduction

Testing methods are models of reality and, therefore, uncertain.

The list of uncertainties of testing methods starts with the well-recognized inter-species and intra-species uncertainty, as well as exposure time uncertainty, adverse effect level uncertainty, high to low dose extrapolation and exposure route uncertainty but extends to uncertainty from the reproducibility of testing results, interpreting complex study results, definition of species-specific mechanisms of action, effects not (yet) detectable with standard tests, endpoint relevance uncertainty, and potential mixture effects. Selection of appropriate statistical methods and significance (p-) values is another critical source of uncertainty. A principle uncertainty of testing methods stems also from the usually pragmatically defined number of animals, defining the sensitivity of a method and thereby what can be considered as a statistically significant effect (see also, e.g., Hartung, 2008). Finally, ignorance needs to be accepted, that is, uncertainty we are not yet aware of and therefore cannot even name (see Fig. 1).

The question is, what do we know about these uncertainties and what do we want to know? The answer to these questions is important in several regards:

First, to develop a more realistic and more transparent hazard assessment, i.e., threshold derivation and adverse effects description, as well as consequent risk assessment, i.e., comparison of threshold levels with exposure estimates is essential. It is important that the uncertainty attached to the hazard and risk estimate be transparent, since this should allow better informed decisions on information and data requirements: if the latter focuses primarily on the aspects of highest uncertainty a positive 3R impact can be gained.

Second, improved transparency for the uncertainty of risk estimates also should allow better informed and, consequently, more responsible risk management decisions, i.e., definition of acceptable exposure levels, acceptable risk, and necessary risk mitigation measures. Transparency for uncertainty may prove essential, especially if such management decisions should, in parallel, be informed by an evaluation of the socioeconomic and environmental sustainability.

Third, transparency for the uncertainty of *in vivo* reference test data is important for validating and accepting new *in vitro* testing methods and non-testing methods (e.g., QSARs). For the calculation of the reliability and relevance of any new method, the reliability and relevance of the reference data is critical if the reference data are used as direct input data for the model development. Moreover, it would be inappropriate, or, in statistical terms, "over-fitting" if *in vitro* or QSAR models were developed to predict the animal testing result with higher accuracy than the animal test's reproducibility or the animal test's expected predictivity of human hazard. In other words, acceptable predictivity of the new methods may be defined by the performance of the actual standard *in vivo* methods. But

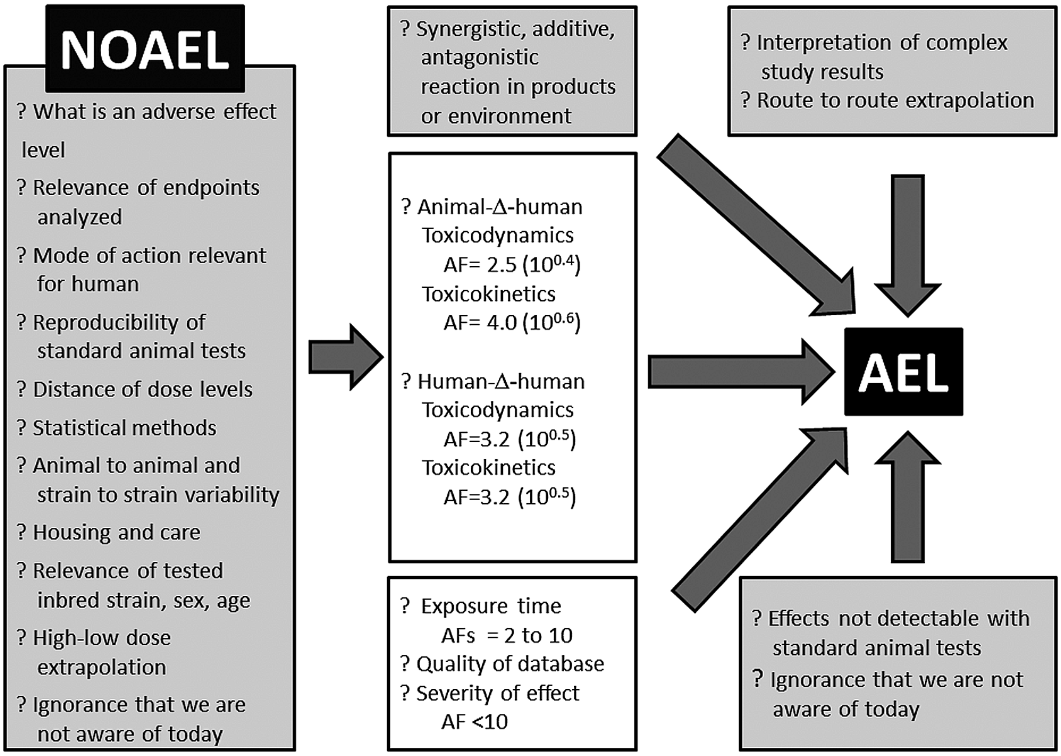

NOAEL
? What is an adverse effect level
? Relevance of endpoints analyzed
? Mode of action relevant for human
? Reproducibility of standard animal tests
? Distance of dose levels
? Statistical methods
? Animal to animal and strain to strain variability
? Housing and care
? Relevance of tested inbred strain, sex, age
? High-low dose extrapolation
? Ignorance that we are not aware of today
? Synergistic, additive, antagonistic reaction in products or environment
? Animal-Δ-human
Toxicodynamics
AF= 2.5 (10^0.4)
Toxicokinetics
AF= 4.0 (10^0.6)
? Human-Δ-human
Toxicodynamics
AF=3.2 (10^0.5)
Toxicokinetics
AF=3.2 (10^0.5)
? Exposure time
AFs = 2 to 10
? Quality of database
? Severity of effect
AF <10
? Interpretation of complex study results
? Route to route extrapolation
AEL
? Effects not detectable with standard animal tests
? Ignorance that we are not aware of today

Fig. 1: Uncertainties of AEL derivation
Several uncertainties should be considered for AEL derivation (grey and white boxes). Usually only inter-species, intra-species, exposure time extrapolation, quality of database, and severity of effects (white boxes) are explicitly considered with quantitative AFs: NOAEL / AF = AEL. NOAEL, No observed adverse effect level; AEL, acceptable exposure level; AF, assessment factor.

innovation and change in the field of testing methods also may be fostered just by scrutinizing the feeling of certainty with data from actual standard methods in quantitative and qualitative terms.

2 What do we know about the uncertainties of testing methods in quantitative terms and how do we usually compensate for them?

Usually the following uncertainties are considered and addressed with standard assessment factors (AFs): inter-species, intra-species, exposure time, and no-observed-adverse-effect-level (NOAEL) uncertainty.

AFs are used by modern (regulatory) safety sciences for the determination of concentrations or doses of particular substances that are acceptable, i.e., pose no or an acceptable minimal threat to the health of exposed individuals. For this purpose, experimental databases of animal or human data are used for the determination of the No Observed Adverse Effect Level (NOAEL) and for the Lowest Observed Adverse Effect Level (LOAEL). The AFs are involved to convert NOAELs and LOAELs to a relevant acceptable exposure level (AEL): AEL = (NOAEL or LOAEL)/AF. AELs are named differently in different regulatory frameworks, e.g., Tolerable Daily Intake (TDI), Acceptable Daily Intake (ADI), Acute Reference Dose (ARfD), Acceptable Operator Exposure Level (AOEL), and sometimes AFs are used to define a Reference Margin of Safety (MOS_{ref}) or Reference Margin of Exposure (MOE_{ref}). However, the principle is the same for all these. If the aim is not only the definition of AELs but estimation of risk, these AELs are compared to measured or estimated exposure levels. If exposure levels are below the AEL, the situation is usually considered safe; if the exposure levels are above the AEL further risk management action usually is required. Various default AFs are published and used in the various regulatory fields (see, e.g., Falk-Filipsson et al., 2007; ECHA, 2012, R.8, Table 8-19), but only recently have the AFs been subjected to scientific analysis in order to develop data-based AFs. In the following we do not discuss the default AFs but focus on available knowledge about uncertainty and thus on the data basis of the more recently proposed data-based AFs.

2.1 Inter-species uncertainty

WHO/IPCS proposed data-based assessment factors for inter-species uncertainty (WHO, 2005): The inter-species uncertainty factor of 10 should be subdivided into a toxicokinetic part and a toxicodynamic part. The first is $10^{0.6}$ (4) for toxicokinetic rat-human extrapolation, and the latter is $10^{0.4}$ (2.5) for toxicodynamic differences. The WHO document indicates that the rat-human toxicokinetic uncertainty factor of 4 is data-based with reference to rat-human differences in "basic physiological parameters that are major determinants of clearance and elimination of chemicals, such as cardiac output and renal and liver blood flows." However, the toxicodynamic inter-species uncertainty factor of 2.5 is not data-based; it represents the remaining to the pragmatic inter-species uncertainty factor of 10. The later published ECHA guidance (ECHA, 2012, R.8, p24) agrees with the concept of allometric scaling (except for substances with primarily local effects or bile excretion or high acute toxicity) and suggests that the factor of 4 for rat-human kinetic differences may be modified to 7 for mouse-human, 2.4 for rabbit-human, and 1.4 for dog-human extrapolations. This is supported by data-based knowledge that smaller animals show faster clearance rates per kilogram body weight compared to larger animals. In the ECHA guidance references to probabilistic assessment, factors are presented that were published after WHO (2005), with the latest being Schneider et al. (2005) (ECHA, 2012, R.8, p70f). This reference and an even later one, Bokkers and Slob (2007) present effect data from identical substances in various species and indicate that the median of distributions of interspecies (NOAEL, LOAEL, or MTD) effect ratios over various numbers of chemicals correspond well with the allometric scaling factors. However, the higher percentiles of these distributions are significantly above the standard inter-species uncertainty factor of 10 (see Tab. 1).

In addition, within the review of Bokkers and Slob (2007) about 23% of the 880 data sets available for sub-chronic toxicity appear inconsistent between mouse and rat (no-response vs. response, increasing vs. decreasing response), even when known rat-specific effects were not taken into consideration. These 23%, plus data sets not usable for other reasons, were excluded from the calculation of the benchmark-dose derived inter-species assessment factor, which indicates that the data-based assessment factor cannot account for all inter-species differences. This finding was not further developed by Bokkers and Slob (2007), (see Tab. 2).

2.2 Intra-species uncertainty

With regard to human intra-species differences, the WHO (2005) document supports a human intra-species uncertainty factor of 10 that should be equally subdivided into $10^{0.5}$ (3.16) for kinetics and $10^{0.5}$ (3.16) for dynamics. This is based on references to Renwick (1993) and Renwick and Lazarus (1998) that summarized effect data for 60 chemicals over various study group sizes (range 5 to 192 subjects); the human intra-species uncertainty factor of 10 covers the 98 percentile. Also for this uncertainty the ECHA guidance references probabilistic assessment factors that were published after WHO (2005), with the latest being Schneider et al. (2005). It reports a refined assessment: Factors were presented for the combination of percentile of substances and percentile of human individuals protected, e.g., indicating a factor of 44 that would protect 95% of adult healthy human individuals for 95% of the substances.

Hasegawa et al. (2010) provided an intra-species estimate on the basis of rat young/newborn NOAEL and LOAEL ratios (18 industrial chemicals), assuming that the largest part from human intra-species uncertainty stems from differences of newborn to young or adult (considering this is the only life stage that usually is not well covered with the standard tests for prenatal, reproductive, and chronic toxicity). A factor of 5 was presented for the 95th percentile (see Tab. 1).

2.3 Exposure time uncertainty

Data-based factors potentially suitable to compensate for exposure time extrapolation from sub-chronic to chronic animal tests were reviewed by Schneider et al. (2005) (the latest reference in ECHA, 2012, R.8, p70f) and further data are provided by Bokkers and Slob (2005) and Batke et al. (2011): 95th percentile values range from 4.7 to 23 (see Tab. 1). However, there is further qualitative uncertainty in these estimates, e.g., within the Bokkers and Slob (2005) publication about 17% of the 314 data sets reported could not be used for the benchmark-dose derived ratio generation due to inconsistent results in the sub-chronic and the chronic study (sub-chronic to chronic: response-scattered, response-no response, increasing-decreasing, decreasing-increasing), 3% showed contradictory results (increasing-decreasing, decreasing-increasing response) and a further 7% of 314 data showed a response in chronic studies, while no response or scattered data were observed in sub-chronic studies (see Tab. 2). Malkiewicz et al. (2009) substantiated that exposure time variability depends on the endpoints analyzed and that mortality related endpoints show a smaller variability compared to non-lethal endpoints. Batke et al. (2011) also reduced several sources of variability, such as dose spacing, study comparability, and target organ comparability (Tab. 1). Pohl et al. (2010) present an approach to refining time extrapolation factors by QSAR-based grouping of substances. Average and geometric mean for acute/intermediate and intermediate/chronic LOAEL ratios are presented for the subgroups volatile organic compounds (VOCs)-aromatic hydrocarbons, VOCs-chlorinated aliphatic hydrocarbons, VOC-chlorinated alkenes, organochlorine pesticides, and organophosphate pesticides. For each subgroup the values are presented separately for the oral and inhalation routes. The presented geometric mean values are considerably lower compared to the other two publications indicated above, but the latter also are derived from many more data pairs.

Besides the extrapolation between sub-acute, sub-chronic, and chronic study designs, there is also uncertainty regarding which exposure study design is relevant for the specific real-life exposure situation, from acute exposure through accidents to life-time exposure with environmental contaminants for pre-natal, post-natal, juvenile, adult, or elderly humans. Test animals differ from humans and environmental organisms not only in their maximal life span but also in their specific pre- and

Tab. 1: Uncertainties usually considered for AF derivation, examples of recent quantitative estimates

Uncertainties	Estimates	Data basis for estimates	Reference
Animal-human inter-species differences – toxicokinetic and toxicodynamic	AS^1 x (GM 1; GSD 3.2) [4.35 (P90); 6.67 (P95); 14.9 (P99)] e.g., for rat-human P95 AF~27	distribution of animal/human dose (MTD) ratios: 63 antineoplastic drugs, 6 species including humans	Schneider et al., 2005, 2006
	AS^2 x (GM 1; GSD 3.4) [8.3 (P95)] e.g., for rat-human P95 AF~40	distribution of mouse/rat NOAEL ratios for identical endpoints: 228 data sets; assume that mouse/rat GSD similar to rat/human GSD	Bokkers and Slob, 2007
	AS^2 x (GM 1; GSD 2) [3.1 (P95)] e.g., for rat-human P95 AF~15	distribution of mouse/rat benchmark dose derived ratio for identical endpoints: 463 data sets; assume that mouse/rat GSD similar to rat/human GSD	Bokkers and Slob, 2007
	AS^2 x (GM 1; GSD 3.8)	distribution of rat/rabbit NOAEL ratios from developmental toxicity studies	Janer et al., 2008
Human intra-species differences	for P90 of individuals: GM 1+2.31; GSD 3.57 [12.8 (P90); 19.8 (P95); 45.7 (P99)] for P95 of individuals: GM 1+3.82; GSD 4.34 [26.1 (P90); 43.8 (P95); 117 (P99)] e.g., 95th percentile of adult healthy humans protected for 95th percentile of substances AF~44 for P99 of individuals: GM 1+8.96; GSD 6.45 [98.7 (P90); 193 (P95); 687 (P99)]	98 data pairs from human case reports	Schneider et al., 2005, 2006
	P95 = 10	based on human variability of kinetic and dynamic parameters: ca. 60 chemicals	Renwick and Lazarus, 1998
	Median 3; GSD 1.38 [5(P95)]	NOAELs and LOAELs ratios [newborn to young rats] for 18 industrial chemicals assuming that the largest part from human intra-species uncertainty stems from differences of newborn to young or adult	Hasegawa et al., 2010
Exposure time extrapolation-sub-chronic to chronic	GM 4.39; GSD 1.82 [9.45 (P90); 11.8 (P95); 17.6 (P99)]	25 sub-chronic /chronic dose ratios derived from NOAELs and LOAELs	Schneider et al., 2005, 2006
	GM 1.7; GSD 2.3 [7 (P95)]	189 sub-chronic/chronic benchmark-dose derived ratios	Bokkers and Slob, 2005
	GM 1.5; GSD 5.3 [23 (P95)]	68 sub-chronic/chronic NOAEL ratios	Bokkers and Slob, 2005
	GM 1.4; GSD 2.1 [4.7 (P95)]	58 sub-chronic/chronic NOAELs and LOAELs ratios	Batke et al., 2011

AF = assessment factor

AS^1 = allometric scaling (rounded figures): mouse 7, rat 4, dog 2, monkey (marmoset) 4, monkey (rhesus) 2

AS^2 = allometric scaling: mouse 10.2, rat 5.1, rabbit 2.6, dog 1.7

GM = geometric mean
GSD = geometric standard deviation

P = percentile

Tab. 2: Uncertainties usually not considered for AF derivation, examples of recent estimates

Uncertainties	Estimates	Data basis for estimates	Reference
Precision or reproducibility of sub-chronic study	range = 10	max NOAEL differences: 77 studies, 36 substances uncertainty may stem from different dose placements, different sets of endpoints, different strains and just chance	Janer et al., 2007
of 2 generation study	range = 10	max NOAEL differences: 25 studies, 12 substances uncertainty may stem from different dose placements, different sets of endpoints, different strains and just chance	Janer et al., 2007
of developmental study	GSD = 3.3	distribution of developmental tox study NOAEL ratios from same substances: 76 studies, 52 comparisons, 21 substances: uncertainty may stem from different dose placements, different sets of endpoints, different strains and just chance	Janer et al., 2008
Inter-species differences – qualitatively inconsistent results (values not used for quantitative interspecies estimate)	23%	increasing vs. decreasing, no-response vs. response, even when known rat-specific effects were not taken into consideration; from 880 available data sets: 58 substances, 91 pairs of studies, 2 sexes, 6 endpoints (body weight at necropsy, liver weight – absolute and relative, kidney weight – absolute and relative, erythrocyte count) minus 212 due to insufficient/not evaluable data	Bokkers and Slob, 2007
Exposure time extrapolation – inconsistent results in sub-chronic and chronic study (values not used for quantitative estimate)	17%	subchronic-chronic: increasing-decreasing, decreasing-increasing response, response-scattered, response-no response; from 314 available data sets: 31 substances, 53 pairs of studies x 2 sexes x 3 endpoints (body weight at necropsy, liver weight – absolute and relative) minus 4 due to insufficient data = 314	Bokkers and Slob, 2005

Exposure time extrapolation – response only in chronic study but not in sub-chronic study (values not used for quantitative estimate)	7%	subchronic-chronic: no response-response, scattered-response;
		from 314 available data sets (see above)
			Bokkers and Slob, 2005

post-natal and juvenile developmental phases of organ systems. Despite knowledge about these differences (Hood, 2006), the respective uncertainty usually is not explicit in hazard and risk assessments.

2.4 NOAEL uncertainty

In human toxicology NOAEL uncertainty usually is recognized only if a lowest-observed-adverse-effect-level (LOAEL) is observed and the NOAEL is extrapolated from this effect level. ECHA (2012, R.8, p70f) informs on default uncertainty factors (between 3 and 10, depending on shape of dose-response curve and severity of effect) as well as benchmark-dose (BMD) modeling for LOAEL to NOAEL extrapolation. It is self-evident that BMD is scientifically superior; nevertheless, it is not routinely applied. Where a NOAEL and a LOAEL were derived, however, there is also uncertainty for toxicological limit value derivation due to dose spacing, the variability of the responses between animals within the dose groups, the definition of "adversity" of an effect, and the use of statistical methods supporting definition of the latter. The current most frequently used statistical approaches to define NOAELs/NOAECs and LOAELs/LOAECs ("hypothesis testing") do not allow describing these uncertainties and have been criticized as conceptually inappropriate for providing quantitative estimates for toxicity; rather, they should be used to answer the fundamental question "Is there a toxic effect?" (see, e.g., Landis and Chapman, 2011; Fox et al., 2012). Several guidance documents address the respective problems (e.g., US-EPA, 2012; OECD, 2006; ECHA, 2012, R.8 and R.10) and estimation techniques such as LCx, ECx or BMD are recommended. These are much less dependent on dose spacing and furthermore curve-fitting of exposure-response data, and the calculation of confidence intervals provides measures for the uncertainty of the toxicity estimates. Use of control group data of within- and between-animal-variation may be helpful to support the definition of adverse effects in terms of a critical effect size (CES) for continuous toxicological parameters (Buist et al., 2009). Fostering the use of these ECx and BMD approaches (Bokkers and Slob, 2007; ECHA, 2012, R.8) and archiving and publishing complete exposure-response data (e.g., by an agreed upon spread-sheet in the journal supplement) would represent very helpful steps towards more transparency of test data uncertainty.

3 What more do we want to know about the uncertainties of testing methods?

There are some critical uncertainties within the evaluation of testing results that often are addressed only if there is substance-specific knowledge. However, more general knowledge and ways to integrate these uncertainties in conclusions on testing results and hazard or limit value estimates may prove critical for scientifically robust hazard and risk assessments: reproducibility of toxicological test results, noise, and contradictory data in the databases used for the derivation of data-based assessment factors, uncertainty from interpretation of complex study results, species-specific mechanisms of action, extrapolation between high and low dose exposure, route-to-route extrapolation, endpoint uncertainty and, finally, ignorance.

3.1 ... expectedly in quantitative terms

Development of data-based assessment factors for the reproducibility of toxicological test results is limited by the fact that toxicological studies are seldom repeated with the same protocol for reasons of costs, animal welfare, the difficulties of overriding positive findings, and the risk of producing false positive findings. Some indications of this uncertainty may be derived from Janer et al. (2007), who report that different subchronic rat studies for a given substance show NOAEL differences up to a factor of 10 (77 studies, 36 substances). The same range factor of 10 is reported for two-generation studies (25 studies, 12 substances). For developmental toxicity studies Janer et al. (2008) report a GSD of 3.3 (see Tab. 2). This reproducibility estimate should represent a variability estimate due to, e.g., animal to animal variability, care variability, and other non-controllable influences.

However, Janer et al. (2008) indicate that the reproducibility estimate also may contain noise from different dose placements, different sets of endpoints, and different strains. This means that there is probably a lot of noise in this estimate for reproducibility. Similar noise may be present in the data-based uncertainty factors for time-extrapolation and inter- and intra-species extrapolation. The noise also may be similar for all the data-based uncertainty factors and may represent a significant portion of the uncertainty. Therefore, the data-based uncertainty factors may not be statistically independent (ECETOC, 2010). However, though the noise could be investigated and eventually lead to refined estimates for reproducibility, as well as time-extrapolation and inter- and intra-species uncertainty, the noise, finally, is also uncertainty that needs to be accounted for when evaluating toxicological data. The noise resolution should be presented not only as reproducibility and dose-spacing uncertainty, but also as endpoint uncertainty or strain-uncertainty or whatever the resolution of the noise results in. The "true" set of endpoints, strain, or other specific test conditions cannot be easily defined.

3.2 ... expectedly only in qualitative terms

As already discussed in chapter 2 for the derivation of data-based estimates for animal-human interspecies uncertainty and for exposure time uncertainty, a relevant part of the database was qualitatively inconsistent or contradictory and therefore needed to be discarded. This points to further uncertainty in terms of data reproducibility or qualitative inter-species differences that are so far not accounted for in the derivation of assessment factors (Tab. 2). It may be that some of this uncertainty is contributed by effects of feeding, housing, and care (Verwer et al., 2007). Qualitative metabolic differences between species and specific life stages also may contribute and, so far, the respective knowledge is limited (Saghir et al., 2012). Some of this uncertainty, like reproducibility (see chapter 3.1), might be better addressed in the future, but it may well be that a larger part of this uncertainty cannot be easily described in quantitative terms and, in the future, should be simply and transparently described in qualitative terms.

Also, the interpretation of complex study results may add uncertainty to study findings: which statistical method or model describes the exposure-response data or which p-value is adequate for defining the NOAEL? Was the systemic effect a consequence of a local effect and does this support adapting the assessment factors? In case of contradictory results: Which results should be given more weight for the conclusion? Is there a specific risk for embryotoxicity that justifies an additional uncertainty factor, or is the observed embryotoxicity just a result of severe maternal toxicity? Disagreement of scientists and regulators on these issues is not the cause but the consequence of this complexity-related type of uncertainty.

A couple of toxicological mechanisms are considered to be test species-specific, e.g., the alpha-2-my-globulin-mediated nephropathy in rats, thyroid gland stimulation by enzyme inducers, forestomach tumors, and others (ECETOC, 2006). Uncertainty may stem from the question of whether the observed effect is species-specific, and what evidence is sufficient to draw a conclusion on that (see, e.g., Ruden, 2002)? On the other hand, there may be mechanisms of action detectable only in humans and not in the test species, e.g., the human eye toxicity of methanol. Such information is even rarer.

The extrapolation from high dose experiments to real world low dose exposure contains, in principle, recognized – but usually not transparent – uncertainty. It relates to the use of toxicokinetic studies for the derivation of, usually, oral and dermal absorption rates. These absorption rates may be used on the one hand to correct the critical NOAEL from the experiment for systemic availability in the test animal, and on the other hand to estimate the real world human systemic exposure. Often the kinetic studies were not carried out at identical doses and test conditions, compared to the experimental studies providing the critical NOAEL (e.g., gavage ± bile duct canulation vs. feeding study); often human exposure is far below the doses of the kinetic study. However, absorption rates depend on exposure doses, concentrations, and test conditions. This may lead to uncertainty with regard to human systemic threshold derivation, as well as human systemic exposure estimates, and consequently risk assessment. Absorption, metabolism, and excretion may change with dose, quantitatively and qualitatively, which complicates extrapolation to dose ranges that were not tested (Creton et al., 2012). Carcinogenic findings identified as critical at high experimental doses and the extrapolation uncertainty to respective no- or minimal-adverse-effect levels are even better recognized (Gaylor, 2005). Furthermore, low-dose effects and non-monotonic dose-response relationships are discussed also, especially in the field of endocrine disruption testing (Vandenberg et al., 2012). All these aspects cause uncertainties for the assessment of dose ranges that were not tested.

Often oral toxicity study results are used for the assessment of risk from dermal and respiratory exposure. The principal uncertainty for this approach is well recognized due to the different impact of local effects and due to a potential first pass effect with oral exposure (see e.g., Rennen et al., 2004). In practice, however, such extrapolation is necessary and practically applied due to limitations with regard to animal use, time, and costs.

It is also relevant to ask whether the (*in vivo* clinical, hematological, histo/pathological, developmental ... or *in vitro*) endpoints analyzed are suitable to cover the relevant adverse effects and whether the best sample number and time point of dosing and endpoint analysis have been chosen. The aim of classification is to define inherent substance properties, but it must not be forgotten that this so-called inherent property, e.g., genotoxicity, is in the end defined not only by the endpoints analyzed but also the points in time and the number of animals used. Endpoint uncertainty also may be different for various chemical groups. For example, for skin sensitization sodium-dodecylsulfate (SDS) represents a recognized false positive in the local lymph nodes assay (LLNA), whereas it is recognized as correct negative in the guinea pig maximization test. In other words, we have to ask if the tested chemistry is within the applicability domain of the *in vivo* model. The concept that various chemistries may need different types of tests is, so far, best recognized with *in silico* and *in vitro* methods but also may be extended to *in vivo* methods.

For some effects, it is clear that they are of high concern, but so far are not detectable with standard tests, like respiratory sensitization. For other aspects of high relevance, testing and evaluation strategies are in intensive development, like endocrine disruption or nanotoxicology and mixture toxicity. For still other aspects, like epigenetic effects, their relevance is even less well understood, and the aspect is just translating from the status of ignorance to uncertainty.

These latter aspects are suitable to highlight that, besides uncertainty, we need to consider ignorance – that is, the uncertainty we are not yet aware of: Our toxicological knowledge is continuously evolving, and with new knowledge and experience we will partially close knowledge gaps and reduce uncertainty. But we also expect that we will discover new fields of uncertainty that we are not aware of today (EEA, 2001).

3.3 ... in specific with regard to nanotoxicology

The risk assessment of nanomaterials is another field containing many uncertainties, and hence the need to address those uncertainties is crucial. Screening reports dealing with nanomaterial risk assessment and nano(eco)toxicological issues identified the testing itself, assessing effects from nanomaterial exposure, nanomaterial characterization, and the exposure assessment as the four main areas of scientific uncertainties (Grieger et al., 2009).

Often the assessor has to deal with a less than adequate test protocol, especially regarding the sample preparation and the measurement of the actual exposure concentration in a test during a certain time period. The inadequate testing leads to unconvincing test results, often with low reproducibility and improper dose/response curves. A major source for these problems is the samples’ tendency to agglomerate and aggregate, especially at higher concentrations, although it also depends on the medium used. The concentration in the test medium may differ from the concentration in the test system, as the particles might absorb to cells.

Risk assessment of a nanomaterial is further complicated by the enormous variety of possible nanomaterials: pure particles, particles with an organic or inorganic coating or both. Also, the particle size distribution, crystal form, shape, specific surface area, zeta potential (surface charge), and properties like dustiness, photocatalytic activity, redox potential and radical formation potential, porosity and pore density, and the ability of the particles to aggregate and agglomerate, as well to dissolute can lead to different test results (compare also OECD, 2008). At the moment, read across to “similar” nanomaterials cannot be performed on a regular basis. Often the nanomaterials are not described properly in the test protocol. But even if a satisfying description is included, it is relevant to know which measurement technique and which analyzer was used, as the differences are still big here.

Also problematic is the limited knowledge base regarding exposure and the usability of models that are based on, for example, water solubility. Taking all these uncertainties into account, the use of the same assessment factors as for conventional chemical substances might be of limited value in some areas of risk assessment.

4 How to improve the accountability of uncertainty of testing method results for hazard and risk assessment?

Systematic framing of the specific questions for specific uncertainties and systematic data collection and synthesis will be necessary to improve uncertainty estimates. The Evidence-based Toxicology initiative might provide a frame fostering further improvements (http://www.ebtox.com). Databases like ToxRefDB (US EPA) and REPDOSE (Bitsch et al., 2006) could be valuable sources for such work.

In addition to improving the knowledge base for uncertainty estimates, an improved representation of uncertainties is necessary within hazard and risk assessment. It usually is recommended to differentiate the various types of uncertainty (Verdonck et al., 2007). So far, for toxicological risk assessment the use of two categories is recognized as most helpful: “variability,” which is system inherent and cannot be reduced with further knowledge, and “uncertainty,” which can be reduced with further knowledge (ECHA, 2012, R.19; SCHER, 2011). For example, the estimate for reproducibility of a test method

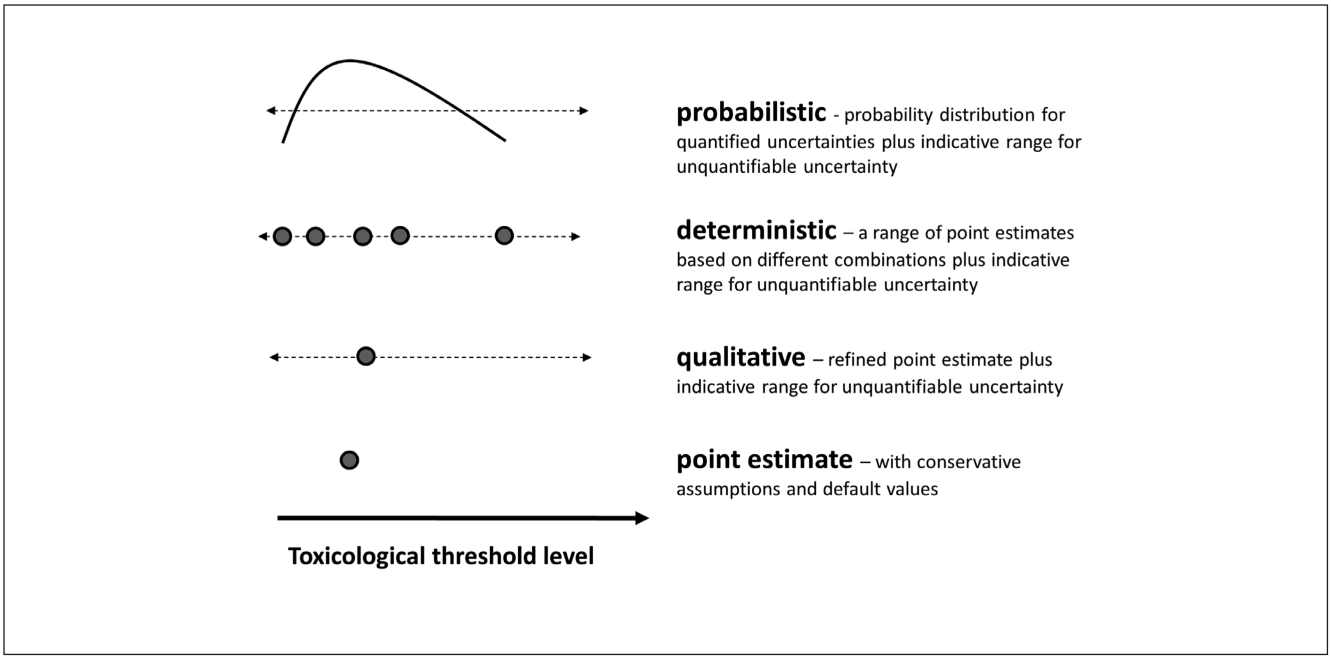
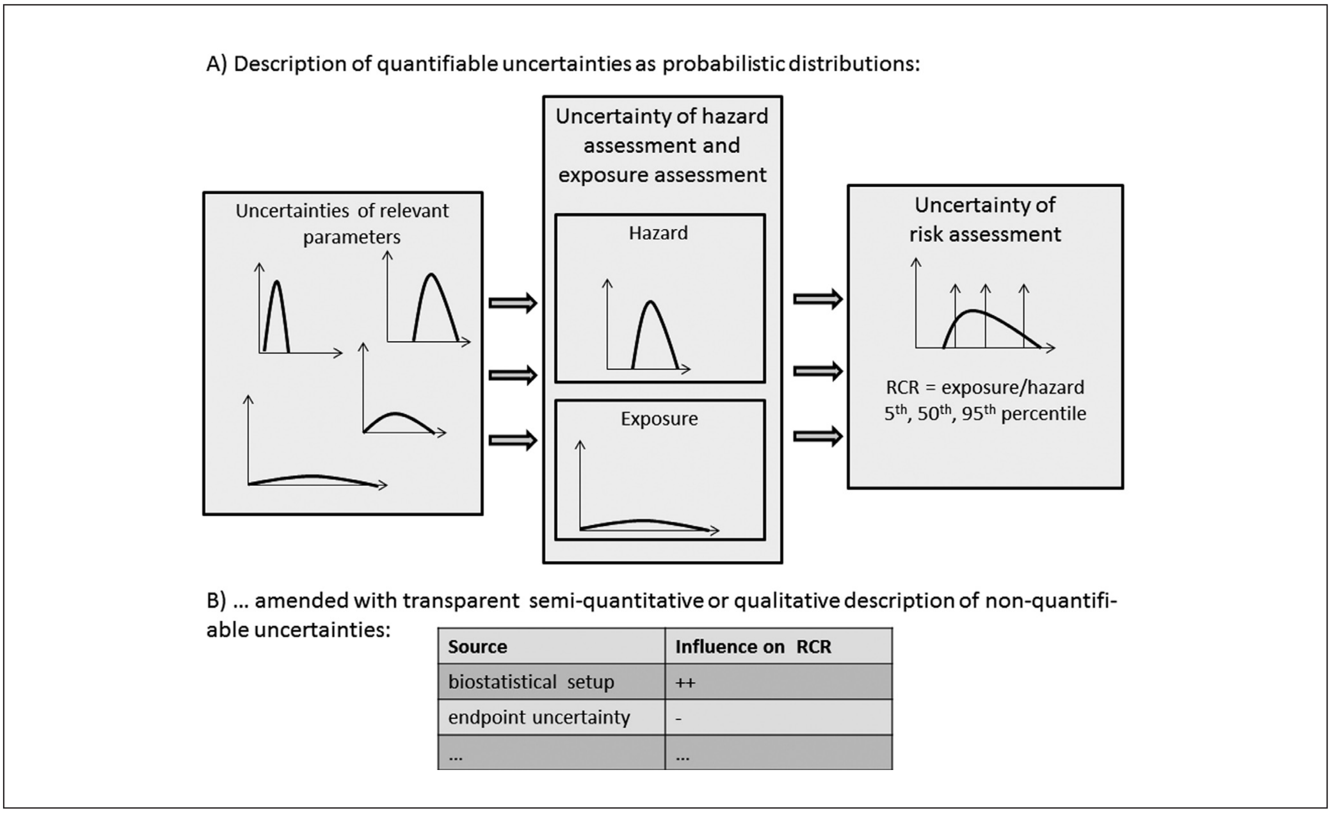

probabilistic - probability distribution for quantified uncertainties plus indicative range for unquantifiable uncertainty
deterministic – a range of point estimates based on different combinations plus indicative range for unquantifiable uncertainty
qualitative – refined point estimate plus indicative range for unquantifiable uncertainty
point estimate – with conservative assumptions and default values
Toxicological threshold level

Fig. 2: Level of uncertainty analysis (modified from ECHA, 2012, R.19)
In the future toxicological threshold levels should not be described as simple point estimates, but should be transparently accomplished with qualitative, deterministic, or probabilistic uncertainty information.

contains "variability," i.e., the animal to animal variability, care variability, and other uncontrollable variability. In addition, the reproducibility estimate contains "uncertainty" in the stricter sense, i.e., noise, for example, from different dose regimens, set of endpoints analyzed, strain differences, and others. Only the latter, the uncertainty, can be reduced with further investigation. Other examples would be the human body weight or metabolism, which are both inherently "variable." Our knowledge of how variable these two aspects are is "uncertain" to different degrees. ECHA (2012, R.19) is more explicit on the characterization of exposure and differentiates between "scenario uncertainty/variability," "model uncertainty/variability," and "input parameter uncertainty/variability." The latter two also are indicated as relevant for hazard characterization and may be used as headers of a checklist sorting various aspects of hazard uncertainty/variability. We propose that the checklist be elaborated further, also taking into consideration the points discussed above in chapter 3, thereby extending the list to uncertainties from all types of standard *in vivo* and *in vitro* tests. Also, the ubiquity of "ignorance" should be recognized. Table 3 may represent a point of departure for the establishment, finally, of case specific lists. So far in this table and this text, differentiation between variability and uncertainty is not explicit, since it is not considered helpful for this overview on testing method uncertainties: All of the listed, (potentially) quantifiable uncertainties include an element of variability and an element of uncertainty, i.e., noise in the databases from which the uncertainty factors were derived. Once quantitatively resolved, the noise should be presented as part of the other variability listed. For purely qualitative uncertainties, differentiation is not considered helpful anyway.

Starting from such a list and depending on the availability of data and information, the uncertainty analysis may be carried out on a qualitative, deterministic, or probabilistic basis. A qualitative assessment may be based on just an uncertainty check-list with verbal explanations and indications of whether and how strong the aspect tends to under- or overestimate the "real" risk with the actual approach or assessment factors applied. For a deterministic assessment, several hazard thresholds may be calculated for various uncertain assumptions, and for a probabilistic assessment more complex bio-statistical modeling is employed to describe the hazard threshold estimate as a distribution covering all known variability and uncertainty aspects. It is acknowledged that ECHA (2012, R.8) and also ECETOC (2010) indicate that simple multiplication of conservative uncertainty factors may lead to inappropriately low limit values, which means that the last approach, i.e., probabilistic integration of uncertainty knowledge appears scientifically superior to the others. The principle of the three approaches was described for the risk ratio (ECHA, 2012, R.19) but can be adapted to describe hazard threshold uncertainty (Fig. 2). The evaluation, however, should allow highlighting the uncertainties with the highest impact on the assessment, so they can be targeted with additional information or data. More elaborated examples of qualitative uncertainty assessment with the use of standard tables in dietary exposure assessments were published by EFSA (2006). SCENIHR (2012) provides a framework for a total weight of evidence evaluation in hazard


A) Description of quantifiable uncertainties as probabilistic distributions:
Uncertainties of relevant parameters
Uncertainty of hazard assessment and exposure assessment
Hazard
Exposure
Uncertainty of risk assessment
RCR = exposure/hazard
5th, 50th, 95th percentile
B) … amended with transparent semi-quantitative or qualitative description of non-quantifiable uncertainties:

Source	Influence on RCR
biostatistical setup	++
endpoint uncertainty	-
…	…

Fig. 3: Description of uncertainties of hazard and risk assessment
A, Description of quantifiable uncertainties as probabilistic distributions; B, amended with transparent semi-quantitative or qualitative description of non-quantifiable uncertainties. Hazard, exposure and risk assessment should be transparent for their uncertainties – in quantitative, semi-quantitative and/or qualitative terms. RCR = risk characterization ratio = exposure/hazard

and risk assessment, including the expression of uncertainties in qualitative and/or quantitative terms, with allowances made to counteract the uncertainties. Other practical examples were published for probabilistic hazard, exposure, and risk assessments (e.g., Bokkers et al., 2009) that integrate inter-species, intra-species and exposure time extrapolation uncertainties and could, in principle, be extended to integrate other quantifiable uncertainties. In any case all aspects of uncertainty that cannot be quantified reliably should be described qualitatively (Fig. 3).

One consequence of this appreciation of uncertainty is that the distinction of so-called threshold effects and non-threshold effects gets blurred; any acceptable exposure level needs to be understood as probability distributed rather than as a sharp cut-off value. Similarly, it may be questioned whether classification with sharp cut-off values, as actually conducted, e.g., for acute toxicity or specific target organ toxicity, should remain the standard approach for the future (see e.g., Hoffmann et al., 2010).

If helpful for specific cases, the ToxR-Tool may be used to judge the reliability of individual testing method results. It is based on the concept of the Klimisch reliability scores and differentiates the various aspects of data reliability in an electronic checklist, allowing a well-targeted discussion (Schneider et al., 2009).

The challenge is not to overburden hazard and risk assessment with uncertainty analysis. Actually, in the context of REACH registrations, uncertainty is barely even addressed. Therefore, some references propose carrying out an uncertainty analysis (for hazard, exposure, and risk) only if the standard approach of risk assessment (based on assessment factors and hazard threshold/exposure risk ratios) indicates that the risk is unacceptable or borderline (ECHA, 2012, R.19; EFSA, 2006). This is justified by the view that the standard risk assessment is carried out with reasonable worst case assumptions and conservative assessment factors. In contrast, others (EEA, 2010) propose that a systematic identification and evaluation of uncertainties should be mandatory for all assessments by industry, regulators, and committees. Some tiered approaches may help maintain a methodology that is proportionate to the needs of the problem. It is considered important to characterize the full range of the potential hazard and risk, including the tails of variability, and make the in/exclusion and weighting criteria transparent.

It also is reported that, further to an appropriate structure for uncertainty evaluation, a proper process is important. Clarity is necessary on who evaluated the uncertainty (e.g., regulators, industry, NGOs, etc.) and how this may influence the outcome of uncertainty assessment. Review by external parties and the possibility of expressing minority opinions may be scientifically important process characteristics as well.

Finally, communication of uncertainty also needs to be improved. It is a difficult and important task, since perceived uncertainty may also negatively impact the quality of life, but it is the improvement of the quality of life that should be the final goal of hazard and risk assessment (EEA, 2010).

As indicated above, principle structures to describe uncertainty are available. However, they need to be filled with further understanding, further data, and freely available practical bio-statistical evaluation tools. They also need to be further tested in practice.

However, describing hazard, and finally risk, in terms of threshold values, risk ratios, and underlying uncertainty may still not be sufficiently helpful for risk management decisions. Scientific committees should be "invited to express opinions on risk in terms of likely impacts on human health and ecosystem services rather than in terms of the more prevalent risk characterizations." This means to express risk in terms of changes of human morbidity or ecosystem services in grading of severity, number of people affected, and probability. It requires exposure consideration, integration of socio-economists, and "training for both assessors and managers based on a common manual." Addressing uncertainty is critical for all aspects of evaluation (SCHER, 2011). This work focuses on testing method uncertainty, and therefore will not elaborate this aspect further. However, it is appreciated that describing hazard uncertainty seems an essential step to allow substantial improvements of the actual hazard and risk assessment outputs, the latter being, so far, not satisfactory for risk management.

5 How to use information on testing uncertainty for defining appropriate prediction models and acceptance criteria for new testing and non-testing approaches?

For *in vitro* test results prediction models are necessary to translate the *in vitro* result to a value or hazard category that can be used for classification and risk assessment. So far, prediction models have been developed on the basis of a training data set, i.e., the data from the reference method that should be predicted, and the data for the new method. Subsequently, the prediction models were validated by a new testing data set that was not used for model development. The approach is, in principle, similar for QSARs development, except that the prediction model input data of the new method are various structural chemical and physicochemical descriptors. Finally, the relevance of the new method is estimated by calculating the correct and false positive, as well as correct and false negative predictions. It is self-explanatory that the reliability, i.e., reproducibility of the reference data, as well as their relevance for predicting human toxicology outcomes, is critical if the reference data are used as direct input data for the prediction model development. It would be inappropriate or in statistical terms, "over-fitting," if *in vitro* or QSAR models were developed to predict the animal testing result with higher accuracy than the animal test's reproducibility or the animal test's expected predictivity of human hazard. This was acknowledged, e.g., within the ACuteTox project for the reproducibility of an acute toxicity category from multiple test results for identical substances (Hoffmann et al., 2010).

However, for deciding on the acceptability of refinement methods like the Extended One Generation Study in rats, knowledge of the uncertainty of the current standard, the two generation rat study, could be helpful. The new method allows the second generation (F2) to be spared, which means a reduction from about 2600 to about 1400 animals, but it maintains the first generation for a longer time period, reduces the pre-mating period, increases the number of animals analyzed (not the number submitted to testing!), and allows testing of a series of new endocrine, as well as developmental neurotoxicity and immunotoxicity endpoints (Piersma et al., 2011; Rorije et al., 2011). It might be easier to accept new approaches if we could have estimates of the reproducibility of the results of the old method, e.g., effects observed in the parent or first generation but not in the second generation, and at least an appreciation in qualitative terms of the uncertainty of the old standard methods with regard to relevance for human toxicity outcomes.

It is also to be expected that new *in vitro* integrated testing approaches, like those envisaged to be based on adverse outcome pathways (AOPs), aim to provide more reproducible and more reliable data compared to the current standard animal tests. They should be mechanistically clear, based on human cell properties, and able to build in human variability. This means that for validation of new human *in vitro* approaches, the scientific rationale may be even more important than their high predictivity of standard animal test data. Practical and ethical aspects may very much support the new approaches, but the key to accepting a change to structurally very different approaches may well be a better understanding of the uncertainties of the new and old methods, quantitatively or just qualitatively.

6 Immediate consequences of the appreciation of testing method uncertainty

It appears that work is necessary to improve the description of uncertainty and allow better informed risk management and data requirement decisions, as well as appropriate prediction models and acceptance criteria for new testing approaches.

However, one immediate consequence of appreciating uncertainty of the current standard testing methods may be to scrutinize the feeling of certainty with the actual testing methods and thereby increase interest in improving hazard characterization with new, more reliable, and more efficient approaches and integrated testing strategies, including *in silico* and *in vitro* methods. The latter may have several advantages

Tab. 3: Improving accountabilty for sources of uncertainty of testing methods[1]

Sources of uncertainty of testing methods		Quantitative estimate[2,3]
Model	**inter-strain and inter-species differences** from anatomy, physiology, molecular targets, ...	
	toxicokinetic	pa
	toxicodynamic	pa
	human intra-species differences from age, size, weight, genetic background, background burden, lifestyle, psychological conditions, disease, ...	
	toxicokinetic	pa
	toxicodynamic	pa
	reproducibility of the test results	
	"bio-statistical chance"	?
	measurement errors	ne
	effects of feeding, housing and care	ne
	bio-statistical setup influencing sensitivity of the model: samples/animals, animals/group, number of groups, dose spacing, ...	?
	background incidences	
	in historical controls	pa
	strain/laboratory specific trends in historical controls	pa
	endpoints analyzed	
	relevant endpoints covered by the model	?
	effects so far not detectable with standard animal tests	ne
	applicability domain	
	specific chemistry leading to false results	?
	evaluation of complex study results	
	validity-criteria, weighting of study findings, biological significance	ne
	use of different statistical models and definition of statistical significance	?
	ignorance (= unknown uncertainties)	ne
Input Parameter	**exposure time extrapolations**	pa
	exposure route extrapolations	?
	exposure dose extrapolation	
	correction from external to systemic dose estimates	pa
	LOAEL to NOAEL	pa
	High-dose (e.g., T25) to low dose (e.g., 10^{-6})	?
	additive, synergistic and antagonistic reactions with other substances (and metabolites thereof) within products, environmental compartments or human body fluids)	?
	ignorance (= unknown uncertainties)	ne

[1] No differentiation between variability and uncertainty is given here, since it is not considered helpful for this overview: All of the listed, (potentially) quantifiable uncertainties include a part of variability that cannot be reduced and a part of uncertainty that could be reduced with further investigation and knowledge, i.e., noise in the databases from which the uncertainty factors were derived. For purely qualitative uncertainties the differentiation is not helpful anyway.

[2] Is a quantitative uncertainty estimate available? pa – at least partially available; ne – not expected; ? – may it be developed for regulatory use?

[3] An additional column should be added to allow a case specific indication in which direction the uncertainty tends to shift the hazard estimate: + …Aspect of uncertainty is likely to overestimate hazard; - Aspect of uncertainty is likely to underestimate hazard; +/- Aspect of uncertainty may over- or underestimate hazard. Eventually this may be extended to a semi-quantitative analysis, e.g., +++ or + or --...

Model and input parameter uncertainties should be summarized to checklists that can be used as points of departure for the establishment of final case specific listings of uncertainties that can be described (semi-) quantively or just qualitatively.

and disadvantages, but at least their uncertainty is, per default, described due to the validation process addressing the applicability domain, reliability, or reproducibility and relevance – issues that are so far not addressed for the old standard animal tests.

Furthermore, it may foster the application of the precautionary principle in the field of risk management (which includes *inter alia*, the requirement for proportional, non-discriminatory, consistent, cost benefit analysis and a clear dedication to scientific review and data improvement (EC, 2000)). It is clearly important to act in time with uncertain data, which is more the standard than the exceptional challenge. It needs to be appreciated that there is a point of data saturation where more data will not reduce uncertainty, and we often do not know if we are in a situation below or above data saturation. More importantly, science often provides a "biodiversity" of evaluations, which reflects the various scientific backgrounds and is natural, human, and consequently also science immanent.

The challenge is to translate the scientific diversity or uncertainty into well-timed political (often yes/no) decisions. With this perception of the task of regulatory scientists, it is appropriate to advocate for precautionary and timely action in terms of acceptable exposure and use definitions. Also, a shift from problem orientation to solution orientation should be fostered, which means *inter alia* to evaluate the risk of certain substance uses in parallel to potential alternatives to these substance uses. Socioeconomic and environmental sustainability analysis may support the 3Rs if carried out early in the testing regime. It may well be that testing and evaluation requirements, as well as the respective uncertainty dilemmas, may be reduced – for well-defined areas – by fostering alternative medicine, alternative pest control, and the rejection of socio-economically easily dispensable products.

7 Conclusion

Recognition of testing method uncertainty appears to be a key for improving scientific hazard and risk assessment, as well as the development of new testing methods and integrated testing strategies.

Quantitative probabilistic uncertainty factors were published for interspecies differences, human intra-species differences, and exposure time extrapolations. The potential to reduce dose spacing uncertainties by benchmark dose approaches for the derivation of acceptable exposure levels is recognized but not routinely used. Some estimates for the reproducibility of testing methods are published, but these estimates could be improved.

Other uncertainties may remain non-quantifiable, like uncertainties stemming from the interpretation of complex study results, qualitative differences in metabolism and physiology between species/strains and life stages, housing and care effects on animals, definition of species-specific mechanisms of action, high dose to low dose extrapolations for systemic exposure and systemic effect estimates, exposure route extrapolation, and endpoint selection uncertainty. Finally, it is reported that ignorance, i.e., uncertainty that cannot even be named since there is not yet awareness of it, should be qualitatively accounted for in hazard and risk assessment. Improvement of the understanding and estimation of testing method uncertainty is highly desirable and should be confronted with a systematic approach that may follow the concept of evidence-based toxicology.

Principle concepts on how to use uncertainty information for hazard and risk characterization have been published, although more elaboration of exposure uncertainty is available. These concepts reflect the need for quantitative uncertainty description, but for several aspects purely qualitative descriptions also are recommended. The latter may be based on checklists that indicate semi-quantitatively, using plus and minus symbols, the overall influence of the uncertainties on the hazard and risk estimates. One consequence of the appreciation of testing method uncertainty is that *any* acceptable exposure level (so-called non-threshold or threshold effects) needs to be understood as probability distributed rather than a sharp cut-off value. In any case, practical application of the available approaches to uncertainty assessment should be fostered, which also would stimulate the relevant methodical improvements.

Improved uncertainty assessment should allow better informed decisions on information and data requirements, since better information and data should be required primarily for the field of highest uncertainty. This practice may support the 3Rs. Furthermore, risk managers could take over more responsibility if they are informed of risk estimates that are more transparent for uncertainty. As an immediate consequence, improved accountability of uncertainty of current standard *in vivo* methods also should foster the development of new *in vitro* and *in silico* approaches: by scrutinizing our feeling of certainty with data from current standard *in vivo* methods, and by characterizing the *in vivo* reference data used for the prediction model development, as well as for the definition of acceptance criteria for the new approaches in terms of reliability and relevance. At the same time, it should foster the application of the precautionary principle for defining acceptable exposure and use. Consequently, an analysis of socioeconomic and environmental sustainability of the regulated products also needs to become a default early in the testing regime.

References

Batke, M., Escher, S., Hoffmann-Doerr, S., et al. (2011). Evaluation of time extrapolation factors based on the database *RepDose*. *Toxicol Lett 205*, 122-129.

Bitsch, A., Jacobi, S., Melber, C., et al. (2006). REPDOSE: A database on repeated dose toxicity studies of commercial chemicals – a multifunctional tool. *Regul Toxicol Pharmacol 46*, 202-210.

Bokkers, B. G. and Slob W. (2005). A comparision of ratio distributions based on the NOAEL and the benchmark approach for subchronic-to-chronic extrapolation. *Toxicol Sci 85*, 1033-1040.

Bokkers, B. G. and Slob W. (2007).Deriving a data-based interspecies assessment factor using the NOAEL and the benchmark dose approach. *Crit Rev Toxicol 37*, 355-373.

Bokkers, B. G., Bakker, M. I., Boon, P. E., et al. (2009). The practicability of the integrated probabilisitc risk assessment (IPRA) approach for substances in food. RIVM report 320121001/2009.

Buist, H. E., Frhr von Bölcshazy, G., Dammann, M., et al. (2009). Derivation of the minimal magnitude of the critical effect size for continuous toxicological parameters from within-animal variation in control group data. *Regul Toxicol Pharmacol 55*, 139-150.

Creton, S., Saghir, S. A., Bartels, M. J., et al. (2012). Use of toxicokinetics to support chemical evaluation: Informing high dose selection and study interpretation. *Regul Toxicol Pharmacol 62*, 241-247.

EC – European Commission (2000). Communication from the commission on the precautionary principle. COM(2000) 1.

ECETOC (2006). Toxicological modes of action: Relevance for human risk assessment. Technical Report 99. http://www.ecetoc.org

ECETOC (2010). Guidance on assessment factors to derive a DNEL. Technical Report No 110. http://www.ecetoc.org

ECHA (2012). Guidance on information requirements and chemical safety assessment, Chapters R.8, R.10, and R.19. http://echa.europa.eu/guidance-documents/guidance-on-reach.

EEA (2001). Late lessons from early warnings: the precautionary principle 1896-2000. http://www.eea.europa.eu/publications/environmental_issue_report_2001_22

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

EFSA (2006). Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment. *The EFSA Journal 438*, 1-54.

Falk-Filipsson, A., Hanaberg, A., Victorin, K., et al. (2007). Assessment factors-applications in health risk assessment of chemicals. *Environ Res 104*, 108-127.

Fox, D. R., Billoir, E., and Charles, S. (2012). What to do with NOACs/NOAELs – prohibition or innovation? *Integr Environ Assess and Manag 8*, 764-766.

Gaylor, D. W. (2005). Are tumor incidence rates from chronic bioassays telling us whatwe need to know about carcinogens? *Regul Toxicol Pharmacol 41*, 128-133.

Grieger, K. D., Hansen, S. F., and Baun A. (2009). The known unknowns of nanomaterials: Describing and characterizing. *Nanotoxicology 2009*, 1-12.

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

Hasegawa, R., Hirata-Koizumi, M., Dourson, M. L., et al. (2010). Proposal of new uncertainty factor application to derive tolerable daily intake. *Regul Toxicol Pharmacol 58*, 237-242.

Hoffmann, S., Kinsner-Ovaskainen, A., Prieto, P., et al. (2010). Acute oral toxicity: Variability, reliability, relevance and interspecies comparison of rodent LD_{50} data from literature surveyed for the ACuteTox project. *Regul Toxicol Pharmacol 58*, 385-407.

Hood R. D. (ed.) (2006). *Developmental and Reproductive Toxicology. A practical approach*. CRC press, Taylor & Francis Group. ISBN 978-0-8493-125-0.

Janer, G., Hakkert, B. C., Piersma, A. H., et al. (2007). A retrospective analysis of the added value of the rat two generation reproductive toxicity study versus the rat subchronic toxicity study. *Reprod Toxicol 24*, 103-113.

Janer, G., Slob, W., Hakkert, B. C., et al. (2008). A retrospective analysis of developmental studies in rat and rabbit: What is the added value of the rabbit as an additional species? *Regul Toxicol Pharmacol 50*, 206-217.

Landis, W. G., and Chapman, P. M. (2011). Well past time to stop using NOAELs and LOAELs. *Integr Environ Assess and Manag 7*, vi-viii.

Malkiewicz, K., Hansson, S. O., and Ruden, C. (2009). Assessment factors for extrapolation from short-term to chronic exposure – are the REACH guidelines adequate? *Toxicol Lett 190*, 16-22.

OECD (2006). Current approaches in the statistical analysis of ecotoxicity data: A Guidance to Application. *OECD Series on Testing and Assessment 54*, EBV/JM/MONO(2006)18.

OECD (2008). List of manufactured nanomaterials and list of endpoints for phase one of the OECD testing programme. *Series on the Safety of Manufactured Nanomaterials 6*, ENV/JM/MONO(2008)13/REV.

Piersma, A. H., Rorije, E., Beekhuijzen, M. E. W., et al. (2011). Combined retrospective analysis of 498 rat multi-generation reproductive toxicity studies: On the impact of parameters related to F1 mating and F2 offspring. *Reprod Toxicol 31*, 392-401.

Pohl, H. R., Chou, C. H., Ruiz, P., and Holler, J. S. (2010). Chemical risk assessment and uncertainty associated with extrapolation across exposure duration. *Regul Toxicol Pharmacol 57*, 18-23.

Rennen, M. A. J., Bouwman, T., Wilschut, A., et al. (2004). Oral-to-inhalation route extrapolation in occupational health risk assessment: a critical assessment. *Regul Toxicol Pharmacol 39*, 5-11.

Renwick, A. G. (1993). Data-derived safety factors for the evaluation of food additives and environmental contaminants. *Food Addit and Contam 10*, 275-305.

Renwick, A. G. and Lazarus, N. R. (1998). Human variability and non-cancer risk assessment – an analysis of the default uncertainty factor. *Regul Toxicol Pharmacol 27*, 3-20.

Rorije, E., Muller, A., Beekhuijzen, M. E., et al. (2011). On the impact of second generation mating and offspring in multi-generation reproductive toxicity studies on classification and labelling of substances in Europe. *Regul Toxicol Pharmacol 61*, 251-260.

Ruden C. (2002). The use of mechanistic data and the handling of scientific uncertainty in carcinogen risk assessments. *Regul Toxicol Pharmacol 35*, 80-94.

Saghir, S. A., Khan, S. A., and Mc Coy, A. T. (2012). Ontogeny of mammalian metabolizing enzymes in humans and animals used in toxicological studies. *Crit Rev Toxicol 42*, 323-357.

SCENIHR (2012). Memorandum on the use of scientific literature forhuman health risk assessment purposes – weighing of evidence and expression of uncertainty. http://ec.europe.eu/health/scientific_committees/policy/index_en.htm

SCHER (2011). Improvement of risk assessment in view of the needs of risk managers and policy makers. http://ec.europa.eu/health/scientific_committees/environmental_risks/index_en.htm

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

Schneider, K., Schuhmacher-Wolz, U., Hassauer, M., et al. (2006). A probabilistic effect assessment model for hazardous substances at the workplace. *Regul Toxicol Pharmacol 44*, 172-181.

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.

US-EPA (2012). Benchmark Dose Technical Guidance. http://www.epa.gov/raf/publications/benchmarkdose.htm

Verdonck, F. A. M., Souren, A., van Asselt, M. B. A., et al. (2007). Improved uncertainty analysis in the European Union risk assessment of chemicals. *Integr Environ Assess and Manag 3*, 333-343.

Verwer, C. M., van der Ven, L. T. M., van den Bos, R., and Hendriksen, C. F. M. (2007). Effects of housing condition on experimental outcome in a reproduction toxicity study. *Regul Toxicol Pharmacol 48*, 184-193.

Vandenberg, L. N., Colborn, T., Hayes, T. B., et al. (2012). Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. *Endocr Rev 33*, 378-455.

WHO (2005). Harmonisation Project Number 2. Chemical-specific adjustment factors forinterspecies differences and human variability: Guidance document for use of data in dose/concentration-response assessment. http://www.who.int/ipcs/publications/en/index.html

Correspondence to

Martin Paparella, PhD
Umweltbundesamt GmbH
Spittelauer Lände 5
1090 Vienna
Austria
e-mail: martin.paparella@umweltbundesamt.at
Phone: +43 1 313 04 3407