"The only certainty is uncertainty."
Pliny the Elder

t4 Workshop Report*

The Probable Future of Toxicology – Probabilistic Risk Assessment

Alexandra Maertens[1], Eric Antignac[2], Emilio Benfenati[3], Denise Bloch[4], Ellen Fritsche[5], Sebastian Hoffmann[6], Joanna Jaworska[7], George Loizou[8], Kevin McNally[8], Przemyslaw Piechota[1], Erwin L. Roggen[9], Marc Teunis[10] and Thomas Hartung[1,11]

[1]Johns Hopkins University, Bloomberg School of Public Health and Whiting School of Engineering, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA; [2]L'Oréal, Research & Innovation, Clichy, France; [3]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy; [4]Department of Pesticides Safety, German Federal Institute for Risk Assessment (BfR), Berlin, Germany; [5]IUF-Leibniz Research Institute for Environmental Medicine, Duesseldorf, Germany; [6]seh consulting + services, Paderborn, Germany; [7]Procter & Gamble, Brussels Innovation Center, Brussels, Belgium; [8]HSE Science and Research Centre, Harpur Hill, Buxton, UK; [9]3Rs Management and Consulting ApS, Kongens Lyngby, Denmark; [10]Institute for Life Sciences & Chemistry, Dept. Innovative Testing in Life Sciences & Chemistry, Hogeschool Utrecht, Utrecht, The Netherlands; [11]University of Konstanz, CAAT-Europe, Konstanz, Germany

Received October 30, 2023;
Accepted January 8, 2024;
Epub January 12, 2023;

Correspondence:
Thomas Hartung, MD, PhD,
Center for Alternatives to Animal Testing (CAAT),
Johns Hopkins University,
615 N Wolfe St., Baltimore,
MD, 21205, USA
(THartun1@jhu.edu)

ALTEX 41(2), 273-281.
doi:10.14573/altex.2310301

Abstract

Both because of the shortcomings of existing risk assessment methodologies, as well as newly available tools to predict hazard and risk with machine learning approaches, there has been an emerging emphasis on probabilistic risk assessment. Increasingly sophisticated artificial intelligence (AI) models can be applied to a plethora of exposure and hazard data to obtain not only predictions for particular endpoints but also to estimate the uncertainty of the risk assessment outcome. This provides the basis for a shift from deterministic to more probabilistic approaches but comes at the cost of an increased complexity of the process as it requires more resources and human expertise. There are still challenges to overcome before a probabilistic paradigm is fully embraced by regulators. Based on an earlier white paper (Maertens et al., 2022), a workshop discussed the prospects, challenges, and path forward for implementing such AI-based probabilistic hazard assessment. Moving forward, we will see the transition from categorized into probabilistic and dose-dependent hazard outcomes, the application of internal thresholds of toxicological concern for data-poor substances, the acknowledgement of user-friendly open-source software, a rise in the expertise of toxicologists required to understand and interpret artificial intelligence models, and the honest communication of uncertainty in risk assessment to the public.

Plain language summary

This workshop report discusses the future of toxicology and how probabilistic risk assessment can help address uncertainties in assessing chemical risks. Experts emphasize the importance of quantitative assessment in toxicology and the need for a deeper understanding of how chemicals affect our health. By incorporating probabilistic risk assessment, we can better evaluate the potential risks posed by chemicals and make more informed decisions to protect human health and the environment. Embracing new technologies like artificial intelligence and natural language processing can enhance data analysis and improve the accuracy of risk assessments in toxicology.

1 Introduction

Toxicology owes its origin to Paracelsus' insight that hazard is not just a property of a chemical but requires a quantitative assessment – there is an amount below which a substance is not a poison. This concept was further refined by Francesco Redi, who realized that the route of exposure (oral vs injected) can mean the difference between life and death after exposure to snake venom (Schickore, 2010). Since then, our understanding of the complexity of both aspects has expanded greatly. To the assumption of a threshold of toxicity was added our understanding that there are both linear and nonlinear dose-response curves; the era of

* A report of t4 – the transatlantic think tank for toxicology, a collaboration of the toxicologically oriented chairs in Baltimore and Konstanz sponsored by the Doerenkamp-Zbinden Foundation. The spectrum of views expressed in this article are those of the contributing authors and do not necessarily reflect those of their institution of employment.

molecular endocrinology demonstrated the importance of non-monotonic dose-response curves. Similarly, our understanding of exposure now encompasses appreciation of the temporal pattern of exposure, the need to understand sensitive populations, and the developmental origins of adult diseases. We have begun to address the importance of multiple co-existing exposures and the interaction of chemical and non-chemical stressors (Sillé et al., 2020), although toxicology has only recently begun to grapple with this.

Human health risk assessment should be considered inherently probabilistic as the risk is a probability for a hazard to occur depending on the exposure. Nonetheless, despite our growing appreciation of the complexity of both dose-response and exposure parameters, risk assessment is still done in a largely deterministic way – with heavy reliance on averages for environmental concentrations, body weight, and intake to produce a point estimate for risk. Moreover, deterministic approaches around the two pillars of risk assessment, namely hazard identification and exposure assessment, along with the allocation of numerous uncertainty factors during the risk characterization step lead to an overly conservative risk estimate. While this enables low-cost first-tier risk assessment for data-poor chemicals and exposure scenarios, a deterministic-based risk assessment does not consider the full range of possible outcomes, nor can it quantify the likelihood of each of these outcomes. Crucial information is lost when risk is reduced to a threshold or hazard to a classification.

The limitations of relying on an average are well-known: Consider the long-running joke about the statistician who drowns in a river with an "average" depth of 3 ft. The joke about a toxicologist might be that, looking at a river with an average depth of 3 ft, they added an uncertainty factor of 10 for geological variability and an additional 10 for database uncertainty, thus assumed an estimated depth of 300 ft and declared the river uncrossable. (Of course, a hazard-based assessor might have assumed the water was no threat at all: It had a high LD_{50} (median lethal dose), was a non-sensitizer, and had no CMR (carcinogenic, mutagenic or reprotoxic) risk. On the other hand, a screening level risk assessor might look at the volume of water in the river compared to the LD_{50} and decide the margin of exposure was unacceptable). The above example is, of course, a simplification and exaggeration. However, it is evident that deterministic risk assessment can be hugely impacted by overestimated exposure, uncertainty factors, and default assumptions. This layering of conservatism can directly lead us to unrealistic decision-making rather than scientifically sound risk assessment. Recently there is a debate regarding the preferable approach for risk assessment, which can either focus on hazard-driven decisions, which are faster and easier, or focus on the identification of a risk, which should require a deeper evaluation, ideally through a probabilistic approach. Another example is the discussion about the use of a mixture assessment factor (MAF), a default hazard factor for each chemical in order to cover "pragmatically" the risk linked to unintentional chemical mixtures (European Commission, 2020; with further comments from the *German Bundesinstitut für Risikobewertung* (Herzler et al., 2021)).

In the real world, there are many examples where over-reliance on simplistic estimates of averages has had detrimental effects – one often-cited case is that of the city of Orange County, CA which in 1994 structured its financial portfolio on the assumption that the then low average interest rates would continue – neglecting both the uncertainty and volatility that interest rates had displayed in the past. Had they made this uncertainty more explicit, they would have realized there was a small but not zero chance that rising interest rates would cost them over 1 billion dollars; interest rates did rise, and the city was forced into bankruptcy (Savage and Markowitz, 2009). The reliance on an average hid enormous tail risks. Indeed, this problem was so acute in the financial world that Monte Carlo simulations were developed specifically to calculate financial risk in a more probabilistic way – an innovation that won the inventor, Harry Markowitz, the Nobel prize in economics[1].

Within toxicology, a probabilistic risk assessment generates a range of risk values, in addition to an average estimate. This approach also seeks to quantify the uncertainty and identify sources of variability. Therefore, a probabilistic risk assessment can identify where more data is needed before a decision can be made confidently, analyze what would be a tipping point, where a decision might be different if assumptions were different, and help to estimate trade-offs with more realistic assumptions. Unlike approaches which require a specific data point – for example, a 90-day NOAEL (no-observed-adverse-effect level) – a probabilistic risk assessment can incorporate multiple types of data. It makes more explicit not just what is known, but what is unknown and uncertain – which in the field of toxicology (and biology in general) encompasses a great deal. Additionally, it provides a more realistic assessment of variability and susceptibility differences. Finally, a probabilistic risk approach lends itself more readily to approaches that build upon big data/artificial intelligence (AI) approaches.

Since the middle of the '90s, probabilistic approaches have been included in the risk assessment paradigm. Thus, thresholds of toxicological concern (TTC; Munro et al., 1996) and benchmark doses (BMD; EFSA Scientific Committee et al., 2019) have gained regulatory acceptance. The field of exposure assessment is probably the area where the most significant progress has been made. Hence, exposure to cosmetic ingredients and products us-

[1] https://www.nobelprize.org/prizes/economic-sciences/1990/press-release/ (accessed 09.12.2022)

Abbreviations: AI, artificial intelligence; AO, adverse outcome; AOP, adverse outcome pathway; BER, bioactivity exposure ratio; BMD, benchmark doses; ChatGPT, Chat Generative Pre-Trained Transformer; DES, diethylstilbestrol; iTTC, internal TTC; MIE, molecular initiating event; NLP, natural language processing; NOAEL, no-observed-adverse-effect level; OS-CAM, Open Science Career Assessment Matrix; PBPK, physiologically based pharmacokinetic modeling; qAOP, quantitative AOP; TL, transfer learning; TTC, thresholds of toxicological concern

ing probabilistic modelling has been accepted for regulatory purposes allowing a more realistic risk assessment[2] (McNamara et al., 2007). More recently, the Crème RIFM probabilistic model approach was successfully shown to be a very robust and useful tool to realistically estimate the aggregate exposure of fragrance ingredients in cosmetic products (Safford et al., 2017).

A robust estimate of the quantity of products and therefore ingredients used by consumers is a crucial step of exposure assessment and consequently risk assessment. Among the key advantages of probabilistic exposure modelling, we may highlight the fact that product usage data are based on real habits without making an assumption that each consumer uses every single product type every day. In summary, these models avoid overly conservative and unrealistic assumptions for estimating exposure and allow a more accurate and realistic risk assessment.

To be sure, toxicology has been well served by the precautionary principle, but there is also a cost to caution. Risk assessments are especially prone to be overly conservative when they are based on multiple high-end point values – the cumulative effect of combining multiple upper-bound assumptions can cause a "compounding of conservatism" and lead to implausible estimates of health effects that fall well above the 90th or 95th percentile of the distribution (Ruffle et al., 2018). Several analyses that focused on Superfund sites found that deterministic and probabilistic risk assessments often had one or more orders of magnitude difference (Viscusi et al., 1997). Given the expense of mitigation, deterministic risk assessments that achieve only a marginal risk reduction for comparatively high social, environmental, and economic costs, are neither practical nor desirable (Ruffle et al., 2018). In this respect, food allergy is a very interesting example. Food allergy is a crucial health concern worldwide, and both the risk assessment and risk management of food allergens have always been a crucial topic/challenge both for industry and regulatory bodies. It is now well established that probabilistic risk assessment is considered the most appropriate method for population allergen risk assessment and management purposes by numerous regulatory bodies (Crevel et al., 2014; Houben et al., 2020).

As a society we face numerous ecological challenges going forward, and many will require complex solutions with novel technologies that will challenge traditional risk assessment paradigms – think of nanotoxicology (Krug et al., 2018) and synthetic biology (Voigt, 2020). A probabilistic risk assessment provides a more precise framework for weighing the advantages and disadvantages of different choices, and this approach can help avoid both regrettable substitutions and burden shifting. How can this be achieved?

2 Data

The field of probabilistic risk assessment developed from the field of engineering and focused on three principal questions: What can go wrong? What are the consequences if something goes wrong? And what is the frequency of these undesirable consequences?

Translating this to toxicology, the first two questions – what can go wrong, and what are the consequences – form the broad topic of chemical hazard and in many respects the foundation of modern toxicology. In some instances, the sequence of events, i.e., the adverse outcome pathway (AOP), is relatively straightforward – a chemical can covalently bind to DNA, this can lead to the adverse outcome (AO) of cancer, or a chemical can covalently bind to a protein and eventually cause skin sensitization. Ideally, we should be able to specify the frequency for these interaction events. Yet even in cases where the molecular initiating event (MIE) is understood and the AO easy to identify, some difficulties emerge. In the case of cancer, very few chemical exposures are linked to unique cancers in a way that can be confidently called causal – for example, prenatal exposure to diethylstilbestrol (DES) causes clear cell vaginal carcinoma (Hoover et al., 2011) and vinyl chloride causes liver angiosarcoma (Sherman, 2009). For most chemical exposures associated with cancer (for example, ethanol and breast cancer (McDonald et al., 2013)) there is no single MIE, and the AO is a shift in the probability of cancer, which may be very slight.

In many areas of toxicology most steps in the toxicological process are not certain – for example, there is no consensus on the precise MIE for most endocrine disruptors, but instead diverse proposals of potential receptor interactions and cellular perturbations (Maertens et al., 2021, 2018). Nor is there any agreement on adverse outcomes, which have been proposed to include delayed puberty, weight gain, but also weight loss, and no consensus on the likelihood. Many areas of toxicology – for example, neurodevelopmental disorders – typically gauge adverse events only at the population level, e.g., a shift in the IQ distribution of the exposed population (Heidari et al., 2022). The existing approaches are also poorly suited to understanding events that are rare, such as birth defects. Valproic acid – often used as a positive control in teratogenicity studies – causes major congenital anomalies in humans with a range of 5 to 15% depending on the dose, and establishing the true frequency requires large, case-controlled population-based studies (Jentink et al., 2010). For chemicals with a less pronounced effect, or an effect observed primarily in a susceptible subpopulation, traditional animal-based studies with the necessary power would likely be too large to be practical and have the added drawback that they are mechanism-blind. Finally, many of the concerns that will likely be critical to risk assessments in the future – for example, the developmental origins of disease, and the unique concerns brought up by co-exposures/mixture toxicology – will require substantial investment in novel methodologies.

Probabilistic risk assessment will therefore require a more mechanistically oriented toxicology, both in terms of MIEs and in terms of understanding the molecular nature of susceptibility – in essence an AOP approach. However, most existing

[2] https://health.ec.europa.eu/publications/sccs-notes-guidance-testing-cosmetic-ingredients-and-their-safety-evaluation-11th-revision_en (accessed 08.06.2023)

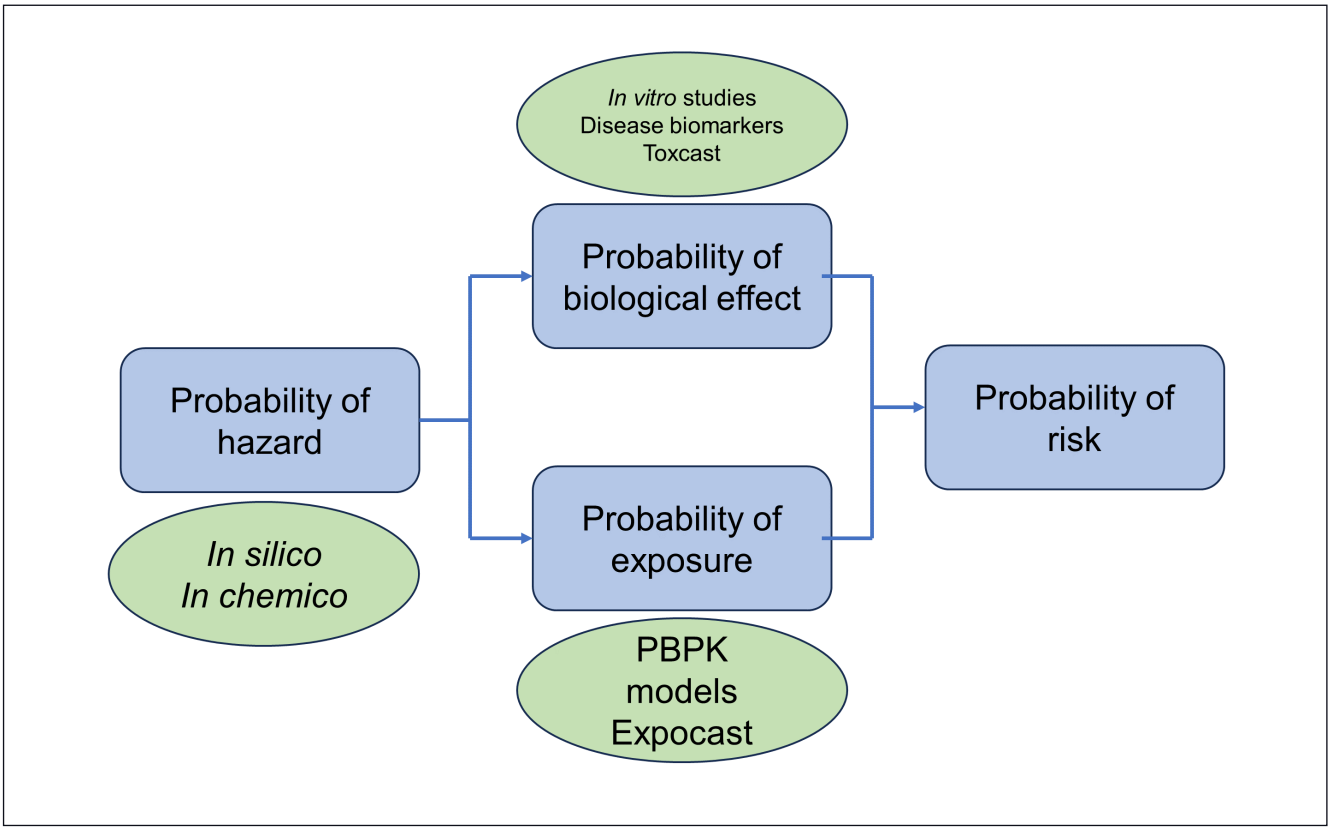

Fig. 1: The probability of risk is ultimately driven by the probability of hazard, probability of exposure, and probability of biological effect

Each of these can be predicted or tested with different data streams, all of which should be considered for predicting risk.

AOPs are qualitative. This is useful to quickly rule out hazards when screening chemicals, but it cannot be used for a quantitative or probabilistic risk assessment of chemicals. Instead, this will require quantitative AOPs (qAOPs), which mathematically define the internal concentrations that trigger the biological tipping points along the AOP, and the probability or magnitude with which those tipping points are exceeded (Spinu et al., 2020). In many respects, a probabilistic risk assessment will be a driving force for 21st century toxicology that avails itself of *in vitro*, *in chemico*, and *in silico* methodologies to map chemical structure to hazard in a more precise way than has been done in the past (Krewski et al., 2010).

On the exposure side, there is an equal need to improve data gathering so that we can base estimates on more extensive exposure data that is both higher in quality and granularity. Remote sensing technologies can improve our ability to monitor external concentrations. Improvements in the technology to miniaturize sensors can allow more ubiquitous sensing of personal exposures, while non-invasive biosampling techniques can improve measurements of internal exposures (Di Guardo et al., 2018). A significantly larger investment in sensing could improve not only the precision of exposure estimates but provide better resolution for temporal and geographic variability. Lastly, in an age with increasingly quantified information about individuals' movements (courtesy of smartphones), and more comprehensive machine-readable electronic health records, our ability to gather data on outcomes provides a far richer, if more high-dimensional, data stream than was possible in the past.

Figure 1 illustrates that the probability of risk is driven by the probabilities of hazard, exposure, and biological effect.

3 Conceptual model

The conceptual model that guides risk assessment – that "hazard × exposure = risk" – remains true for a probabilistic risk assessment with some refinements. While hazard is still often handled with binary labels (e.g., sensitizer or non-sensitizer, carcinogen or non-carcinogen), or relatively simple binning strategies (e.g., low, medium, or high acute toxicity), hazard will instead need to be quantified as a point-of-departure, BMD, or potency with a confidence interval.

One existing methodology, the TTC can be viewed as probabilistic, as it was based on a cumulative distribution of oral NOAELs. Despite having been developed on the basis of a small data set, it has proven remarkably robust (Kroes et al., 2005). In

an era of large data sets, it has the potential to be further refined with closer mapping of functional groups to toxicity, and therefore can be extended beyond its current uses in food additives (Hartung, 2017). In addition, as our ability to extrapolate from external exposures to internal exposures improves, using physiologically based pharmacokinetic modeling (PBPK) or *in-vitro*-to-*in-vivo* extrapolation (IVIVE), we can focus more precisely on internal dose. This can be extended to an internal TTC (iTTC) (Ellison et al., 2021) – basically, a refinement of the TTC concept based on plasma concentrations for the risk assessment of substances with a low absorption (either by the oral or the dermal route), as the internal exposure is in these cases more relevant than the external exposure. Accordingly, an interim iTTC of 1 µM was proposed by Blackburn et al. (2019) based on experiences from the pharmaceutical industry, an in-depth review of published non-drug chemical/receptor interactions, and an analysis of ToxCast™ data. In addition, chemicals excluded from the interim iTTC approach comprise the original TTC exclusions, such as androgen and estrogen agonists. Interestingly, Najjar et al. (2023) recently published a case study that provides a practical example of how the iTTC can be used to refine a TTC-based assessment for dermal exposures to consumer products and provide a promising tool for risk assessment when systemic exposure data are available. The authors compared Cmax values measured in clinical pharmacokinetic studies performed with seven chemicals topically applied in over-conservative conditions to the interim 1 µM iTTC value. Although they indicated that refinements and improvements were still needed, the comparison of Cmax values derived from PBPK with the iTTC interim value could become a very potent probabilistic model for a pragmatic risk assessment approach.

Deterministic approaches, which compare estimates of exposures (often based on simplistic metrics such as production volume) to a 90-day NOAEL for an appropriate margin of exposure, will instead look at a probability of exposure that, based on PBPK data, is tied to a probability of tissue concentration. In lieu of a margin of exposure, therefore, we have a comparison of a probability distribution for both tissue concentrations and an iTTC or point of departure from a qAOP. This avoids the difficulty of interspecies comparison as well as makes more explicit physiological variation in ADME (adsorption, distribution, metabolism and excretion). This approach is also more easily extendible to the problem of multiple chemical stressors, as it can focus on instances where there is a similar target organ or overlap in AOPs.

Where both exposure and PBPK modelling indicate limited tissue concentration of a chemical with minimal potential for hazard, we can confidently declare there is a low probability of risk. For much of the middle ground, we have an area of uncertainty which will likely prompt testing to clarify and refine the estimates, as well as areas where the risk is easy to identify. This allows risk assessors to better take advantage of data streams, ranging from ExpoCast to Toxcast (Houck et al., 2013) for exposure and hazard as well as using approaches such as RASAR (read-across structure activity relationships) for potency (Luechtefeld et al., 2018). In lieu of replacing uncertainty with conservatism, at each step an assessor can pinpoint an area of uncertainty and seek to gather more data if necessary, as well as ask if the models are too poorly parameterized, or if the model is too complex, such that a cascade of uncertainty makes a decision impossible.

The cosmetics industry is a very good example of how a risk assessment can shift from deterministic to probabilistic approaches. During the last two-three decades, the cosmetic industry had to deal with a regulatory animal testing ban in Europe, which encompassed first finished products and then ingredients. This industry developed some pioneering approaches to tackle the consequences of the regulatory ban, and it demonstrated the capacity to operationally handle this topic. The risk assessment approach aimed to protect consumers as published by Middleton et al. (2022) provides a full probabilistic model of risk assessment, since it combines not only probabilistic exposure assessment by using PBPK modelling, but also bioactivities and points of departure derived using probabilistic approaches to determine a bioactivity exposure ratio (BER), i.e., a margin of safety. Numerous case studies using this modelling approach have been performed and published providing refinements and improvements. This model could become the first end-to-end probabilistic risk assessment model, though the road to gaining regulatory acceptance may still be long.

4 Resources

A probabilistic risk assessment will necessarily be more computationally complicated than a deterministic risk assessment, and therefore computational infrastructure will be required. This means not just creating the necessary software, but also acknowledging that software creation is a scientific contribution of equal value to a publication. This leads to the premise that funding agencies will need to adapt their funding programs and evaluation procedures to include the support of robust and reproducible research outcomes beyond scientific publications in high impact journals. The EU Commission has already started the discussion on this transition and aims to implement the evaluation of scientists and organizations using, e.g., an Open Science Career Assessment Matrix (OS-CAM), which illustrates the range of evaluation criteria for assessing Open Science activities (EC et al., 2017). Funding support must also be provided not just for software creation but also for its maintenance, and equally crucial, to improve user experience so that it is accessible to the broadest range of users. Currently, many of the most sophisticated models typically require a user to be comfortable with a command-line interface as well as having the experience and expertise to trouble-shoot software installation, which can often require complicated dependencies. Some areas of science such as the field of computational oncology long ago realized that powerful models, with appropriate attention to a user interface and data visualization, could be made broadly accessible (Gao et al., 2014).

The creation of additional software and models for performing probabilistic risk assessments of a diverse nature, brings with it the challenge of standardization, e.g., in the area of omics tech-

nologies. In the early days of transcriptomics data analysis from array technology there was little standardization, which led to inconsistencies in scientific outcomes between experiments and different labs. This has changed dramatically with the adoption of open-source programming languages in this field of research. Nowadays, Bioconductor[3] is the leading platform for data analysis of many high-throughput technologies, including genomics, metabolomics, and proteomics (Gentleman et al., 2004). Similarly, the last decade also saw the rise of publicly available curated bioactivity databases such as ChEMBL or ToxCast. These resources have chemoinformatics interfaces and allow, for instance, performing chemical similarity searches; to achieve that, some level of chemical structure curation and standardization must be provided. The open-source data science-oriented programming languages such as Python and R have greatly changed the way analysis can be made reproducible. Using version control software like Git and platforms such as Github[4] has helped the re-use and adoption of existing software tools, and adaptation of the code to solving new problems in many scientific fields, including toxicology (Peng and Hicks, 2021).

Above and beyond producing the crucial kind of mechanistic data to inform probabilistic risk assessment, it is equally important to have scientific data that is machine-readable (Luechtefeld et al., 2018). This will almost certainly require some level of enforcement mechanism from funding agencies as well as journals, in addition to a commitment to database infrastructure. While virtually everyone in the community wants both free and well-maintained, well-documented software as well as easily accessible, curated, publicly available data, that priority is not currently reflected in funding mechanisms – a mismatch of community needs and resources that will have to be addressed.

As data sources continue to grow, gathering and parsing data is no longer feasible to be done by one person or even a committee. Both the FDA and the EPA have embraced the necessity of using AI approaches for literature review for both risk assessments and systematic reviews, and probabilistic risk assessment will certainly require a similar approach. To leverage AI for data extraction, accumulation, and synthesis from a large volume of (unstructured) data sources, we need to examine the technology of natural language processing (NLP). Recently, the field of language generative AI has made a major leap forward with the release of ChatGPT[5] (Chat Generative Pre-Trained Transformer) by OpenAI. The success of ChatGPT encouraged other researchers to develop more domain-specific language models such as BioGPT, which can be used for mining biomedical content (Luo et al., 2022). These approaches could be further utilized for constructing (q)AOPs, which is currently a very labor-intensive and time-consuming process. Workflows that leverage NLP to develop AOPs have been recently proposed (Corradi et al., 2022). Indeed, as probabilistic risk assessment will be heavily dependent on AI (and the role of AI will only grow over time), this will require a commitment to AI that is both transparent and robust.

It is apparent that the advances in AI and the availability of relevant Big Data will greatly enhance the development of probabilistic approaches to risk assessment. However, with computationally intensive solutions, the demand for powerful computational hardware also increases. There seems to be an increasing interest and incentive for funding agencies to allocate parts of their budgets to fund applications from researchers to use computational infrastructure. Research organizations have been investing in building human capital with the know-how on leveraging cloud-infrastructure for computation. Large hardware facilities already exist, and some of them are even international collaborations as exemplified by the EU High Performance GRID[6].

Finally, we should not forget the human resources. The use of NAMs, application of or development of more advanced computational models, and development and maintenance of software will require large and very diverse groups of experts. It is likely that these groups of researchers at organizations will have to adapt new ways of working to move from a more classical "content-driven" scientific composition to a more agile team with flexible expertise and working disciplines. These agile teams often consist of specific researchers who have in-depth knowledge on, e.g., immunology or cancer biology, mixed with expertise on IT-related topics such as cloud-computing, DevOps, software development, or web development. Often, these teams will also have data science and statistical expertise on board and have the reproducible research principle (including FAIR principles) in scope when delivering their results to the scientific community (Baxter et al., 2022).

5 Capacity building

Currently, risk assessment has many different paradigms depending on whether a substance is a drug, cosmetic, an industrial chemical, or a food additive. Drug regulators understand that some risk is inherent in medicine and therefore focus on risk vs benefit. Risk assessors for industrial chemicals typically assume chronic exposure and must pay more attention to the diversity of exposure scenarios as well as accidental exposures, but rarely think of risk trade-offs. Those who focus on food additives must manage a great deal of chemical diversity and often focus on a TTC, with the understanding that absolute safety is impossible. Cosmetic risk assessors must concern themselves with potential contaminants as well as consumer perceptions about safety and testing. Often, however, scientists trained in one risk assessment paradigm rarely look to other areas for ideas, and they view the chemical space of their focus as unique and outside the applicability domain of models or systems developed in other areas. However, AI and transfer learning (TL) thrive on large data sets

3 http://new.bioconductor.org/

4 https://github.com/

5 https://chat.openai.com/auth/login

6 https://www.e-cam2020.eu/pilot-projects-with-industry/ (accessed 08.06.2023)

that explore a broad chemical space. A probabilistic risk assessment framework is by necessity a multi-disciplinary undertaking and therefore will require individuals with different types of expertise – this will mean a re-orienting towards team science as well as better communication amongst regulators, industry, and academia. Since a probabilistic risk assessment will likely require a fundamental retooling of paradigms, this opens space for better communication amongst agencies. In many agencies where probabilistic risk assessment is performed, such as EFSA, EPA, and the OECD, case studies can be used to highlight the methodologies and the advantages.

Finally, there will be little point in developing probabilistic risk assessment as a field if regulatory acceptance does not follow. This will require a shift amongst regulators, and potentially, the legal landscape within which they operate, as uncertainty offers an opening for much legal contestation. It will require greater transparency about uncertainty, and more fundamentally, regulators will need to be comfortable both with a more statistically complicated approach and greater reliance on predictive methodologies and AI. Insofar as a probabilistic risk assessment will gather broader types of data, it will require greater scrutiny of data quality, bias in data, and data provenance – not to mention the necessity of using AI to gather data.

6 Education

Probabilistic risk assessment will almost certainly require an increase in statistical literacy of both users and practitioners, and points to a critical need to increase the grounding in statistics in training. In addition, as toxicology switches to a more mechanistic, qAOP focus, this will require a more solid grounding in quantitative biology – in this respect, toxicology will simply be following the pattern of the life sciences in general (Eaton et al., 2020).

Similarly, toxicologists of the future will require literacy in machine learning. Not every toxicologist needs to be an expert in deep learning – in fact most do not – but an understanding of how AI uses data, how to make sense of the outputs, whether as a classification or a summary statistic, how to benchmark AI performance, and how to understand both the strengths and weaknesses of AI approaches should be a skill imparted to every student. Indeed, many of the challenges of AI are not the more esoteric aspects of the algorithms used to build the model but instead more quotidian problems: The data used for the model are flawed, biased or somehow unrepresentative of the real world, or the conclusions and interpretation are unjustified by the result (a problem not unique to AI). Identifying these flaws requires minimal computational expertise.

Toxicologists – even those planning on staying in academia – need to be trained in the regulatory use of data, and more broadly on the critical need for data to be robust. This will include focusing on adequately powered studies, a high degree of reproducibility, and adequate quality control. Many fields have faced a reproducibility crisis, and the life sciences in general and toxicology in particular are no exception (Hartung, 2013; Ioannidis, 2005).

Correlation versus causality needs to be more explicit in our thinking and our education – while virtually all scientists (and most laypeople) will confidently assert that correlation is not causation, the reality is they are often conflated, or ambiguous terms are used (e.g., a chemical is "linked with" an outcome). A study found that nearly one third of papers on obesity used inappropriate causal language in their abstracts, even when study design precluded such conclusions (Cofield et al., 2010). In point of fact, what distinguishes mere association from causation is often a source of discussion in environmental epidemiology (Lucas and McMichael, 2005). As large-scale data becomes more common, AI and big data-oriented approaches will find connections that would otherwise be missed. At the same time, there is the potential for the field to be overwhelmed by spurious correlations. AI does not reduce the need for critical thinking – if anything, it sharpens it.

7 Communication

Probabilistic risk assessment will require a substantial investment in communication – end users will have to be able to understand the outcome without statistical jargon. The field of medicine has long understood that probability is often counterintuitive and difficult to grasp for many practitioners (Arkes et al., 2022), and realized that presenting data on false positives and negatives for a test did little to impact clinical practice, while presenting data "on number needed to test" was more easily understood and implemented. For risk assessment, reframing the output as the fraction of people protected is an easy-to-understand metric for decision-making.

The reality is that the existing hazard-based classification system that underpins many aspects of chemical regulation will likely persist and needs to be accommodated within the new framework. The output for many risk managers in the field may therefore remain the same, while additional data will be available on an as-needed basis. A probabilistic risk assessment should not seek to start anew, but rather build upon this framework to add nuance where necessary while keeping simplicity where possible.

8 Conclusion

Significant progress has been made over the last decades regarding the use of probabilistic modelling in risk assessment. The continuous development of NAMs both in the field of hazard identification and exposure assessment will without doubt provide us with robust and pragmatic tools to conduct realistic risk assessment and, above all, to ensure the safety of consumers.

However, beyond the need to continue to develop robust NAMs, in the field of learning and education there should be an effort to change this mindset and avoid over-conservatism. The way to address uncertainty is at the heart of the debate (Rusyn and Chiu, 2022; Dourson et al., 2022).

Mel Andersen, one of the main proponents of transforming toxicology towards a more mechanistic, systems biology-based approach, was once asked how many "pathways of toxicity" can we expect to find? He responded succinctly "*132*" and then added "*as a toxicologist/risk assessor, I am accustomed to false*

accuracy" (Kleensang et al., 2014). Toxicology, as a field, has been challenged with providing assurance of safety in the presence of enormous uncertainty. In the future, instead of looking away from areas where there is uncertainty, we must instead seek to shine a light on them. Moreover, we cannot pretend to have more certainty than we actually have – the illusion of accuracy is often more problematic than an acknowledgement of where uncertainty exists. Finally, perfect safety will always be an illusion – attempts to eliminate risk are often simply shifting risk from more visible to less visible areas. Embracing the intrinsic uncertainty in science and the inevitability of risk is essential for a 21st century toxicology.

References

Arkes, H. R., Aberegg, S. K. and Arpin, K. A. (2022). Analysis of physicians' probability estimates of a medical outcome based on a sequence of events. *JAMA Netw Open 5*, e2218804. doi:10.1001/jamanetworkopen.2022.18804

Baxter, A. L., BenZvi, S. Y., Bonivento, W. et al. (2022). Collaborative experience between scientific software projects using agile scrum development. *Softw Pract Exp 52*, 2077-2096. doi:10.1002/spe.3120

Blackburn, K. L., Ellison, C. A., Stuard, S. B. et al. (2019). Dosimetry considerations for in vivo and in vitro test data and a novel surrogate ITTC approach for read-across based on metabolites. *Comput Toxicol 10*, 145-157. doi:10.1016/j.comtox.2018.08.005

Cofield, S. S., Corona, R. V. and Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obes Facts 3*, 353-356. doi:10.1159/000322940

Corradi, M. P. F., de Haan, A. M., Staumont, B. et al. (2022). Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies. *Biomater Biosyst 7*, 100061. doi:10.1016/j.bbiosy.2022.100061

Crevel, R. W., Baumert, J. L., Baka, A. et al. (2014). Development and evolution of risk assessment for food allergens. *Food Chem Toxicol 67*, 262-276. doi:10.1016/j.fct.2014.01.032

Di Guardo, A., Gouin, T., MacLeod, M. et al. (2018). EnvironG mental fate and exposure models: Advances and challenges in 21st century chemical risk assessment. *Environ Sci 20*, 58-71. doi:10.1039/c7em00568g

Dourson, M., Ewart, L., Fitzpatrick, S. C. et al. (2022). The future of uncertainty factors with in vitro studies using human cells. *Toxicol Sci 186*, 12-17. doi:10.1093/toxsci/kfab134

Eaton, C. D., Lamar, M. D. and McCarthy, M. L. (2020). 21st century reform efforts in undergraduate quantitative biology education: Conversations, initiatives, and curriculum change in the United States of America. *Lett Biomath* 7, 55. https://scarab.bates.edu/faculty_publications/345/

EFSA Scientific Committee, More, S. J., Bampidis, V. et al. (2019). Guidance on the use of the threshold of toxicological concern approach in food safety assessment. *EFSA J 17*, e05708. doi:10.2903/j.efsa.2019.5708

Ellison, C. A., Api, A. M., Becker, R. A. et al. (2021). Internal threshold of toxicological concern (iTTC): Where we are today and what is possible in the near future. *Front Toxicol 2*, 621541. doi:10.3389/ftox.2020.621541

EC – European Commission, Directorate-General for Research and Innovation, Valdes, C. et al. (2017). Evaluation of research careers fully acknowledging Open Science practices: Rewards, incentives and/or recognition for researchers practicing Open Science. *Publications Office of the European Union*. https://data.europa.eu/doi/10.2777/75255

European Commission (2020). Chemicals Strategy for Sustainability – Towards a toxic-free environment communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM/2020/667 final. European Commission, Brussels, Belgium. https://eur-lex.europa.eu/legal-content/en/txt/pdf/?uri=celex:52020dc0667

Gao, J., Aksoy, B. A., Gross, B. et al. (2014). Abstract 4271: The CBioPortal for cancer genomics as a clinical decision support tool. *Cancer Res 74, Suppl*, 4271. doi:10.1158/1538-7445.am2014-4271

Gentleman, R. C., Carey, V. J., Bates, D. M. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol 5*, R80. doi:10.1186/gb-2004-5-10-r80

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

Hartung, T. (2017). Thresholds of toxicological concern – Setting a threshold for testing below which there is little concern. *ALTEX 34*, 331-351. doi:10.14573/altex.1707011

Heidari, S., Mostafaei, S., Razazian, N. et al. (2022). The effect of lead exposure on IQ test scores in children under 12 years: A systematic review and meta-analysis of case-control studies. *Syst Rev 11*, 106. doi:10.1186/s13643-022-01963-y

Herzler, M., Marx-Stoelting, P., Pirow, R. et al. (2021). The "EU chemicals strategy for sustainability" questions regulatory toxicology as we know it: Is it all rooted in sound scientific evidence? *Arch Toxicol* 95, 2589-2601. doi:10.1007/s00204-021-03091-3

Hoover, R. N., Hyer, M., Pfeiffer, R. M. et al. (2011). Adverse health outcomes in women exposed in utero to diethylstilbestrol. *N Engl J Med 365*, 1304-1314. doi:10.1056/nejmoa1013961

Houben, G. F., Baumert, J. L., Blom, W. M. et al. (2020). Full range of population eliciting dose values for 14 priority allergenic foods and recommendations for use in risk characterization. *Food Chem Toxicol 146*, 111831. doi:10.1016/j.fct.2020.111831

Houck, K. A., Richard, A. M., Judson, R. S. et al. (2013). ToxCast: Predicting toxicity potential through high-throughput bioactivity profiling. In P. Steinberg (ed.), *High-Throughput Screening Methods in Toxicity Testing* (1-31). John Wiley & Sons, Inc. doi:10.1002/9781118538203.ch1

Ioannidis J. P. (2005). Why most published research findings are false. *PLoS Med 2*, e124. doi:10.1371/journal.pmed.0020124

Jentink, J., Loane, M. A., Dolk, H. et al. (2010). Valproic acid monotherapy in pregnancy and major congenital malformations. *N Engl J Med 362*, 2185-2193. doi:10.1056/nejmoa0907328

Kleensang, A., Maertens, A., Rosenberg, M. et al. (2014). Pathways

of toxicity. *ALTEX 31*, 53-61. doi:10.14573/altex.1309261

Krewski, D., Acosta, D., Jr, Andersen, M. et al. (2010). Toxicity testing in the 21st century: A vision and a strategy. *J Toxicol Environ Health B Crit Rev 13*, 51-138. doi:10.1080/10937404.2010.483176

Kroes, R., Kleiner, J. and Renwick, A. (2005). The threshold of toxicological concern concept in risk assessment. *Toxicol Sci 86*, 226-230. doi:10.1093/toxsci/kfi169

Krug, H. F., Bohmer, N., Kühnel, D. et al. (2018). The DaNa2.0 knowledge base nanomaterials – An important measure accompanying nanomaterials development. *Nanomaterials 8*, 204. doi:10.3390/nano8040204

Lucas, R. M. and McMichael, A. J. (2005). Association or causation: Evaluating links between "environment and disease". *Bull World Health Organ 83*, 792-795.

Luechtefeld, T., Marsh, D., Rowlands, C. et al. (2018). Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci 165*, 198-212. doi:10.1093/toxsci/kfy152

Luo, R., Sun, L., Xia, Y. et al. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform 23*, bbac409. doi:10.1093/bib/bbac409

Maertens, A., Tran, V., Kleensang, A. et al. (2018). Weighted gene correlation network analysis (WGCNA) reveals novel transcription factors associated with bisphenol A dose-response. *Front Genet 9*, 508. doi:10.3389/fgene.2018.00508

Maertens, A., Golden, E. and Hartung, T. (2021). Avoiding regrettable substitutions: Green toxicology for sustainable chemistry. *ACS Sustain Chem Eng 9*, 7749-7758. doi:10.1021/acssuschemeng.0c09435

Maertens, A., Golden, E., Luechtefeld, T. H. et al. (2022). Probabiɵlistic risk assessment – The keystone for the future of toxicology. *ALTEX 39*, 3-29. doi:10.14573/altex.2201081

McDonald, J. A., Goyal, A. and Terry, M. B. (2013). Alcohol intake and breast cancer risk: Weighing the overall evidence. *Curr Breast Cancer Rep 5*, 208-221. doi:10.1007/s12609-013-0114-z

McNamara, C., Rohan, D., Golden, D. et al. (2007). Probabilistic modelling of European consumer exposure to cosmetic products. *Food Chem Toxicol 45*, 2086-2096. doi:10.1016/j.fct.2007.06.037

Middleton, A. M., Reynolds, J., Cable, S. et al. (2022). Are non-animal systemic safety assessments protective? A toolbox and workflow. *Toxicol Sci 189*, 124-147. doi:10.1093/toxsci/kfac068

Munro, I. C., Ford, R. A., Kennepohl, E. et al. (1996). Correlation of structural class with no-observed-effect levels: A proposal for establishing a threshold of concern. *Food Chem Toxicol 34*, 829-867. doi:10.1016/s0278-6915(96)00049-x

Najjar, A., Ellison, C. A., Gregoire, S. et al. (2023). Practical application of the interim internal threshold of toxicological concern (iTTC): A case study based on clinical data. *Arch Toxicol 97*, 155-164. doi:10.1007/s00204-022-03371-6

Peng, R. D. and Hicks, S. C. (2021). Reproducible research: A retrospective. *Ann Rev Publ Health 42*, 79-93. doi:10.1146/annurev-publhealth-012420-105110

Ruffle, B., Henderson, J., Murphy-Hagan, C. et al. (2018). Application of probabilistic risk assessment: Evaluating remedial alternatives at the Portland Harbor Superfund Site, Portland, Oregon, USA. *Integr Environ Assess Manag 14*, 63-78. doi:10.1002/ieam.1999

Rusyn, I. and Chiu, W. A. (2022). Decision-making with new approach methodologies: Time to replace default uncertainty factors with data. *Toxicol Sci 189*, 148-149. doi:10.1093/toxsci/kfac033

Safford, B., Api, A. M., Barratt, C. et al. (2017). Application of the expanded Creme RIFM consumer exposure model to fragrance ingredients in cosmetic, personal care and air care products. *Regul Toxicol Pharmacol 86*, 148-156. doi:10.1016/j.yrtph.2017.02.021

Savage, S. L. and Markowitz, H. M. (2009). *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. John Wiley & Sons. https://play.google.com/store/books/details?id=2lsLAQi0LlcC

Schickore, J. (2010). Trying again and again: Multiple repetitions in early modern reports of experiments on snake bites. *Early Sci Med 15*, 567-617. http://www.jstor.org/stable/20787431

Sherman, M. (2009). Vinyl chloride and the liver. *J Hepatology 51*, 1074-1081. doi:10.1016/j.jhep.2009.09.012

Sillé, F. C. M., Karakitsios, S., Kleensang, A. et al. (2020). The exposome – A new approach for risk assessment. *ALTEX 37*, 3-23. doi:10.14573/altex.2001051

Spinu, N., Cronin, M. T. D., Enoch, S. J. et al. (2020). Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. *Arch Toxicol 94*, 1497-1510. doi:10.1007/s00204-020-02774-7

Viscusi, W. K., Hamilton, J. T. and Dockins, P. C. (1997). Conservative versus mean risk assessments: Implications for superfund policies. *J Environ Econ Manage 34*, 187-206. doi:10.1006/jeem.1997.1012

Voigt, C. A. (2020). Synthetic biology 2020-2030: Six commercially-available products that are changing our world. *Nat Commun 11*, 6379. doi:10.1038/s41467-020-20122-2

Conflict of interest

None declared.

Data availability

No datasets were generated in this study.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 963845 (ONTOX). We would like to thank Dr Bas Bokkers, National Institute for Public Health and the Environment (RIVM), The Netherlands, for valuable discussion.