1 (2025) 100020

Contents lists available at ScienceDirect

NAM Journal

journal homepage: www.sciencedirect.com/journal/nam-journal

Full Length Article

Highlighting best practices to advance next-generation risk assessment of cosmetic ingredients

Eric Vaillancourt [a,*], Alaa Afdal Mohamed [b], Jay Ansell [c], Takao Ashikaga [d], Ahmed Ayad [e], Hadeel Ayman [b], Marize Valadares Campos [f], Han-Sheng Chien [g], Arianna Giusti [h], Masato Hatao [i], Frances Hill [j], Abeer Khattab [b], Reinhard Kreiling [k], Ah Young Lee [l], Miao Li [m], Pinpin Lin [n], Yen-Chun Lin [g], Victor Mencarelli [o], Stanley Milstein [m,1], Beta P. Montemayor [p], Carolina Motter Catarino [q], Rawda Negm [b], Gladys Ouédraogo [r], Eun Jae Park [s], Mustafa Varçin [h], Tzu-Yi Yang [g], Yu Janet Zang [m], Matthew Philip Dent [t]

[a] Health Canada, Healthy Environments and Consumer Safety Branch, 269 Laurier Ave. W., Ottawa, Ontario K1A 0K9, Canada
[b] Egyptian Drug Authority (EDA), Egypt
[c] US Personal Care Products Council (PCPC), 555 13th Street NW, WA, DC 20004, USA
[d] National Institute of Health Sciences (NIHS), Japan
[e] Chamber of Pharmaceutical Industries, Cosmetics and Medical Devices, Federation of Egyptian Industry (CPCMD), Egypt
[f] Brazilian Health Regulatory Agency (ANVISA), Gerência-Geral de Cosméticos e Saneantes, Setor de Indústria e Abastecimento (SIA), Trecho 5, Área Especial 57, CEP 71205-050, Brasília, DF, Brazil
[g] Taiwan Food and Drug Administration (TFDA), Chinese Taipei
[h] Cosmetics Europe, The Personal Care Association, Avenue Herrmann Debroux 40, 1160 Brussels, Belgium
[i] Japan Cosmetic Industry Association (JCIA), Metro City Kamiyacho 6F, 5-1-5, Toranomon, Minato-ku, Tokyo 105-0001, Japan
[j] Office for Product Safety and Standards (OPSS), UK
[k] Clariant Produkte (Deutschland) GmbH, Am Unisyspark 1, 65843, Sulzbach, Germany
[l] Korean Cosmetics Association (KCA), Republic of Korea
[m] US Food and Drug Administration (US FDA), 5001 Campus Drive, College Park, MD 20740, USA
[n] National Health Research Institutes, Taiwan
[o] Independent Beauty Association (IBA), United States, 14901 Quorum Drive, Suite #630, Addison, TX 75254, USA
[p] Cosmetics Alliance Canada, 420 Britannia Road East, Suite 102, Mississauga, ON L4Z 3L5, Canada
[q] Association of the Cosmetic, Toiletry and Fragrance Industry (ABIHPEC), Av. Paulista, 1313 Cerqueira César, São Paulo, SP 01311-000, Brazil
[r] L'Oréal, Research and Development, Paris, France
[s] Ministry of Food and Drug Safety (MFDS), Republic of Korea
[t] Unilever Safety, Environmental and Regulatory Science, Sharnbrook, Bedfordshire, MK44 1LQ, UK

ARTICLE INFO

Keywords:
Next-generation risk assessment
New approach methodologies
Cosmetics risk assessment
Non-animal approaches
Cosmetic ingredient safety

ABSTRACT

Public expectations and regulatory demands have driven a paradigm shift in cosmetic ingredient safety assessment. Next-generation risk assessment (NGRA) and new approach methodologies (NAMs) provide an opportunity to implement alternative procedures which are more ethical and biologically-relevant, have the potential to become more cost-effective and time-efficient, and are at least as protective of human health as traditional, animal-based approaches. However, the sheer number and diversity of the rapidly-evolving NAMs currently available can be confusing and overwhelming to safety assessors. To successfully transition from principles to routine application, NAMs must gain acceptance within safety assessment communities. Harmonizing NAMs and establishing standardized risk assessment workflows are paramount to gaining regulatory acceptance. Consequently, the International Cooperation on Cosmetics Regulation (ICCR) has tasked a team of scientists from regulatory authorities and the cosmetic industry to define best practices which align with the ICCR principles of NGRA for cosmetic ingredients, to help advance the acceptance and use of NAMs in cosmetic safety assessment and, where applicable, regulatory decision-making.

* Corresponding author.
E-mail addresses: Eric.Vaillancourt@hc-sc.gc.ca (E. Vaillancourt), Matthew.Dent@unilever.com (M.P. Dent).
[1] Retired.

https://doi.org/10.1016/j.namjnl.2025.100020
Received 18 March 2025; Received in revised form 12 April 2025; Accepted 12 April 2025
Available online 13 April 2025

1. Introduction

Risk assessors have a duty to ensure that the safety assessments they conduct are fit-for-purpose and draw on the most appropriate science. While extensive guidance is available concerning traditional animal-based risk assessment, little guidance exists describing how non-animal approaches should be developed and applied for the safety assessment of cosmetic ingredients. The ICCR principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients (the 'ICCR Principles of next generation risk assessment (NGRA)') were published in 2018 (Dent et al., 2018), and the impact of this paper indicates a high demand among scientists for guidance and support in how to develop and apply new, non-animal approaches to the risk assessment of cosmetic ingredients. The EU's Scientific Committee on Consumer Safety (SCCS) has risen to this challenge by incorporating the use of NGRA in the 12th revision to the Notes of Guidance (SCCS, 2023) for the safety assessment of cosmetics. These Notes of Guidance reference the ICCR Principles of NGRA and describe the general workflow for its application (Berggren et al., 2017). The overall conclusion of both the SCCS and the ICCR is that progress is being made in the development and application of NGRA, but that more examples are needed to increase confidence that NGRA is protective, including for new compounds (Dent et al., 2021; Rogiers et al., 2020).

Such examples of NGRA for cosmetic ingredients are now starting to appear in the literature, ranging from application of next generation read-across (Bury et al., 2021; Ouédraogo et al., 2022), endpoint-specific examples such as skin sensitization (Gilmour et al., 2020; Reynolds et al., 2021; Vandecasteele et al., 2021), liver steatosis (Sepehri et al., 2025), and protective approaches for general systemic toxicity (Baltazar et al., 2020; Ebmeyer et al., 2024). However, while NGRA is developing and the level of confidence with different approaches is still advancing, there are few agreed-upon ways to approach NGRA. OECD Guideline 497 on defined approaches for skin sensitization (OECD, 2021a) is the first internationally harmonized guideline describing the non-animal approaches that can be used to replace animal tests to identify skin sensitizers. Such guidelines do not yet exist for more complex health effects such as systemic or reproductive and developmental toxicity, and there remains a lack of guidance on how to ensure individual safety assessments that use these approaches are robust and scientifically acceptable.

Alongside increasing consumer demand for cosmetic products and ingredients that have not been tested on animals, regulatory bans on animal testing necessitate the use of novel approaches. For example, the EU Cosmetics Regulation (EC) No 1223/2009 led the way in prohibiting the use of animals in cosmetic ingredient safety assessment as well as the marketing of any cosmetic product that relies on animal-derived safety data generated after specific deadlines. The stringent legal framework renders development and application of animal-free approaches essential in the EU and in other geographies imposing such bans, and contrasts with the possibility of using animal testing as a last resort. To reflect this, in this paper the term NAM refers exclusively to non-animal new approach methodologies.

It is also important to note that cosmetics-related NGRA approaches strongly rely on NAMs which may not always have undergone formal regulatory validation. Whereas many of the NAMs used in NGRA are scientifically sound and could be deemed scientifically valid, this scientific validity does not directly translate into regulatory acceptance. As such, case studies currently help advance the field of NGRA, but may not yet represent acceptable methodologies for regulatory decision-making. This difference is particularly important in the areas of systemic toxicity and long-term toxicity, where it is clear that traditional validation which is focussed on a one-to-one replacement concept is not fit-for-purpose, and a new paradigm is needed to facilitate scientific and regulatory acceptance (van der Zalm et al., 2022).

To begin to address the need for guidance for safety assessors, a group of scientists working in cosmetic safety evaluation was convened by the International Cooperation on Cosmetics Regulation (ICCR) to review some currently available examples of using NGRA, and to see how these can inform best practice. The goal of a NGRA is to be human-relevant, exposure-led, hypothesis-driven, and designed to prevent harm to consumers. These first 4 over-riding ICCR principles set the context for the remaining 5 principles, which describe how an NGRA should be conducted (using a tiered and iterative approach, following an appropriate literature search and evaluation of the available data, and using robust and relevant methods and strategies) and how the assessment should be documented (transparent and explicit about the logic of the approach and sources of uncertainty). These 9 ICCR principles work together to ensure the quality and relevance of the safety assessment, and in their practical application, many of them interact and overlap with one another, and can be difficult to separate. For this reason, this report focusses on the 4 principles that describe the goal of the NGRA, and discusses the important aspects of how the assessment should be conducted and documented as they relate to these 4 over-riding principles. This discussion is illustrated with case studies found in the literature, which demonstrate different elements of the ICCR principles and their possible application to safety decision-making.

2. The overall goal is a human safety assessment

Toxicology and other safety sciences developed rapidly during the mid-20th century, when the molecular interactions that can lead to pathological changes were not well understood. This meant that the most practical way to understand the safety of chemicals was to dose laboratory animals at high levels of exposure and observe the outcome. However, as we now understand, there are many examples of toxicities seen in animals that cannot be directly related to effects in humans (Bartsch et al., 2018; Chamanza and Wright, 2015; Cook et al., 1999; Cunha et al., 2015; Leist and Hartung, 2013). Therefore, although animal test data (when used within an appropriate decision-making framework) have generally provided a high level of protection for consumers, the pace at which science has developed in recent years provides a turning point for integrating more human-relevant data to inform safety decisions (Krewski et al., 2010). Furthermore, not only do scientists today have a vast array of non-animal-based methods available to characterise the biological activity of chemicals, but there has also been tremendous progress in understanding the importance of exposure in safety decision-making, and in developing tools to better characterise exposures to both single chemicals and mixtures in a more human-relevant way (Desalegn et al., 2019; OECD, 2004). Safety assessors today have an unprecedented opportunity to capitalise on the recent scientific advances in molecular biology, computational and exposure science to underpin robust safety decisions, rather than inappropriately using *in vitro* or *in silico* methods to seek to solely replicate the results of apical animal studies (Dent et al., 2018).

Common themes or key factors identified that should be considered when approaching this principle in practice, are summarized below. Ensuring these key factors are explicitly covered will increase the robustness of a safety assessment:

1. Consider and document the relevance of the hypothesis/hypotheses that needs/need to be tested.
2. Identify exposures that are relevant to the use context.
3. Identify human-relevant test systems.
4. Identify assays and/or biomarkers that are relevant to the hypothesis being tested, and explain why they were selected.

2.1. Consider and document the relevance of the hypothesis/hypotheses that needs/need to be tested

Practical application of the ICCR principles of NGRA are described later. These reveal the potential intersection and overlap of some of the

principles and best practices covering how to document the hypothesis-driven nature of the NGRA and how to ensure the assessment is exposure-led. However, even at the problem formulation stage when the hypotheses to be tested are being defined and documented, care needs to be taken to ensure these are human-relevant, thus they need to consider the route and extent of exposure to the cosmetic ingredient. The key factors listed above, therefore, need to be considered in parallel rather than sequentially. It is critical to devote time and attention to the problem formulation stage of the safety assessment to prevent this being reduced to a list of standardized tests to perform, and to ensure the scientific reasons for generating the data are not forgotten and replaced by a 'tick-box' mentality. This means that where a data gap exists that prevents a safety decision from being made, assessors need to resist the urge to simply find a way to replicate the animal test without using animals. For example, where the data gap relates to skin sensitization potency, researchers and risk assessors have elucidated the biological events that drive the induction of skin sensitization which must be understood and accounted for before determining which approaches (including exposure-based waiving, (Q)SARs, *in vitro* assays) would be appropriate to provide the necessary information. This is easier when there is a defined adverse outcome pathway (AOP), which is the case for skin sensitization, where several approaches are available in the literature which take an exposure-led and hypothesis-driven approach based on a common safety assessment framework (Gilmour et al., 2020). This framework was based on the SEURAT-1 *ab initio* framework for systemic effects (Berggren et al., 2017), but tailored specifically to address the skin allergy AOP. *"Ab initio"* (literally "from the beginning") refers to performing a safety assessment on a substance with no existing data from first principles. Using this example, in the absence of substance-specific data or the ability to perform a read-across, the assessor would first determine the likelihood that the chemical would be sufficiently potent to induce skin allergy at relevant levels of exposure (Safford, 2008). If a consumer safety risk cannot be ruled out, enough is known of this AOP and the molecular initiating events (MIEs) and key events (KEs) necessary to cause skin sensitization to enable the development of key hypotheses and a weight-of-evidence (WoE) approach to test them. Critically, understanding the biological rationale and human relevance of different data types and of the overall approach facilitates interpretation where different NAMs provide discordant data (Vandecasteele et al., 2021).

This is more challenging where there are data gaps for more complex health effects such as systemic toxicity, which encompasses a multitude of adverse effects and potential AOPs. In these cases, it is simply not feasible to address every possible AOP, but instead to develop hypotheses based on general biological activity (Thomas et al., 2019). One common systemic toxicological data gap is developmental and reproductive toxicity (DART) (Rovida and Hartung, 2009), which does not represent a single AOP. However, even in these cases, the SEURAT-1 *ab initio* framework provides a useful guide to complete an assessment. As per the skin allergy example, the first step would be to determine the likelihood that the chemical is a sufficiently potent developmental or reproductive toxicant to pose a health risk at relevant exposure levels. Various tools are available to help answer this question, including interrogating existing data and the use of structural alerts and additional uncertainty factors (Wu et al., 2013). Where a safety decision cannot be made using such approaches, even without an AOP or AOPs, human-relevant hypotheses still can be formed to guide the data needed to complete the risk assessment by using available knowledge to assess the ability of the chemical to perturb events known to be critical for successful human reproduction and development (Rajagopal et al., 2022).

2.2. Identify exposures that are relevant to the use context

One of the most important ways to ensure human relevance of any toxicological risk assessment is to ensure that the exposure data are relevant to real-world exposure scenarios. Therefore, a best practice NGRA will include exposure estimates from authoritative sources or from high-quality and representative consumer habits and practices surveys. Several examples of safety assessments that exemplify this principle are available in the literature, including systemic safety assessments on face cream or body lotion (Baltazar et al., 2020; OECD, 2021b), and skin allergy safety assessments on body lotion (Vandecasteele et al., 2021). These assessments use published habits and practices survey data (Hall et al., 2007) which are also referenced in the SCCS Notes of Guidance (SCCS, 2023). Exposure assessments typically follow a tiered approach, both in terms of the external (applied) exposures derived from an understanding of consumer habits and practices, and in terms of the internal exposures. This means the exposures considered can range from simple, worst-case deterministic use data to (in extremely rare cases) measured biomonitoring data in a relevant human population. In practical terms, the tool most likely to be used in NGRA to establish internal exposures is physiologically-based kinetic (PBK) modelling as demonstrated in the phenoxyethanol OECD Integrated Approach to Testing and Assessment (IATA) case study (OECD, 2021b). In this case study, population variability in kinetic parameters was incorporated in the PBK modelling; this population-based approach was used to refine the human-relevant internal exposure estimate. The challenge for safety assessors is ensuring the output of the PBK model is indeed human-relevant, because often there is a lack of human measured data to parameterise models or to verify output. It is therefore important for PBK models used in cosmetic safety assessment to be protective (*i.e.*, sufficiently conservative) to cover the exposed populations.

2.3. Identify human-relevant test systems

At the core of the seminal 2007 report from the National Research Council, "Toxicity Testing in the 21st Century – a Vision and a Strategy" was the suggestion that the test systems (cells, cell lines, cellular components) used to provide toxicity information should preferably be of human origin (Krewski et al., 2010). For example, in a 2020 OECD IATA case study that used a NAM-supplemented read-across approach to evaluate the safety of dermal exposure to propylparaben from cosmetics (OECD, 2020a), the assay selected to provide information on estrogenic potency was the ER-Calux® assay. The ER-Calux® cells are a human osteosarcoma cell line (U2-OS) that stably expresses human estrogen receptor alpha (ER1), thus meeting the criteria of providing a human-relevant test system for evaluating effects mediated by ER1 binding. One possible limitation of this study relating to this principle was the use of rat S9 to investigate the effects of metabolism on the ER response, rather than human S9. Due to a mixture of practicality, availability, and reproducibility issues, it is still more common for rat S9 to be used than its human equivalent. However, all other assays and approaches used to characterise the absorption of propylparaben and its biotransformation in both skin and liver were based on human tissues, cells or S9 fractions.

Any test system is by nature a model, so it will never be able to mimic human biology exactly. However, there is an opportunity for a NAM-based approach to represent human biology more closely and, as experience with animal models has shown, it is not necessary to exactly mimic human biology to arrive at a protective safety decision. Many examples of NGRA currently in the literature make use of human cancer cell lines (Baltazar et al., 2020; OECD, 2021b) and yet, for the most part, these appear to provide points of departure that are at least as protective as the animal tests (Middleton et al., 2022; Paul-Friedman et al., 2020). Notwithstanding their advantages, cancer cell lines are clearly not representative of healthy cells in an *in vivo* situation, and higher-tier (primary cells, organotypic models, micro-physiological) systems may need to be deployed on a case-by-case basis depending on the need of the assessment. One limitation in the phenoxyethanol IATA case study (OECD, 2021b) was that, despite the PBK model predicting that the

kidney would be the compartment exposed to the greatest concentration of the major stable metabolite of phenoxyethanol, bioactivity data generated in a kidney-relevant cell model were not provided and, therefore, it was difficult to ensure that perturbations within pathways expressed in such cells were not missed. Therefore, an important part of the documentation of an *in vitro* based assessment is the justification for the choice of cells and culture system, which should consider the extent of exposure of different body compartments to the parent compound and major metabolites.

2.4. Identify assays and biomarkers that are relevant to the hypothesis being tested, and explain why they were selected

Where the hypothesis concerns a defined mode of action, selection of assays and biomarkers may be more straightforward. For example, in the propylparaben IATA mentioned above (OECD, 2020a), the read-across approach was based on comparing the estrogenic potency of the target molecule with that of source molecules (other parabens). The availability of a commercial reporter gene assay that measures ER-mediated estrogenic activity was therefore the obvious choice. However, in cases where there is no defined mode of action, identifying biomarkers for more complex health effects can be challenging. One approach that can be used is to find signatures using methods such as metabolomics or transcriptomics that are predictive of the effect being assessed. This was the method used to identify the altered ratio of ornithine/cysteine (two metabolites in human embryonic stem cell media) as a predictive biomarker for identifying possible human teratogens (Palmer et al., 2013). This metabolic biomarker was later tested in the ToxCast library, where it demonstrated high specificity and modest sensitivity (Zurlinden et al., 2020). Another approach is to use an understanding of the pathways that, if perturbed, can lead to adverse health effects, and to deploy biomarkers that can characterise alterations in those pathways. This method was employed to define a cell stress panel (Hatherell et al., 2020), which included biomarkers relevant to several stress response pathways that, if perturbed, can lead to target organ toxicity. In the case of this cell stress panel, the hypothesis was chemical agnostic, as it focussed on a set of modes of action that could lead to toxicity. There are also examples of substance-specific best practice in the literature, where the hypothesis and, therefore, methods are much more tailored to the mode of action of the chemical being assessed. One of these is an *in vitro* safety assessment on per- and poly-fluorinated substances (PFAS) (Fragki *et al.* 2023). PFAS have been associated with lipid disturbances in epidemiological studies, and some have been shown to be hepatotoxic, as evidenced by increased liver weight, steatosis or necrosis in animal studies. Considering that triglyceride (TG) accumulation within HepaRG cells has been suggested to be a biomarker for hepatosteatosis and hepatotoxicity (Lichtenstein et al., 2020) as well as evidence from gene expression data in PFAS-treated HepaRG cells, Fragki et al. (2023) used TG accumulation as a biomarker for determining a benchmark dose for PFAS in HepaRG cells. Note that whether the approach that is taken is targeted or untargeted, it is crucial to ensure that changes in the biomarkers can usefully distinguish between levels of risk associated with chemical exposure (*i.e.*, to ensure that the approach is relevant and robust).

3. The assessment is exposure-led

During the mid-20th century, hazard-based assessments gained prominence as there were improvements in the understanding of toxicological principles and chemical hazards. They were prevalent until the late 20th century, when exposure-based risk assessments emerged as a complementary, holistic approach which offered a more realistic and practical representation of the actual risk faced by individuals/populations. Exposure-based risk assessments have become increasingly viewed as best practice in recent decades, and this shift can be attributed to several factors including, but not limited to, *i*) the availability of improved data collection and analysis methods and technologies (*e.g.*, biomonitoring techniques, computer modelling) which enable more accurate assessment of exposure levels and pathways, *ii*) an increased regulatory emphasis on risk management that requires the development of targeted and effective strategies to protect human health, and *iii*) public concerns and stakeholder engagement which require increased transparency to provide a clearer understanding of the risks posed by specific hazards. In addition, special consideration should be given to external *vs* internal exposure as well as to various refinement approaches, and examples are summarized hereafter.

3.1. External and internal exposure

Exposure plays a crucial role in the risk assessment process. It involves estimating or measuring the magnitude, frequency, and duration of exposure to a substance, as well as considering the characteristics of the exposed population. In the context of cosmetic ingredient safety assessment, exposure includes both external and internal components, and it determines the relevance of hazard data for risk characterization. In cases of "low" exposure, exposure-based waiving approaches such as the Threshold of Toxicological Concern (TTC) or Dermal Sensitization Threshold (DST) can be considered within a WoE approach. The external part of exposure focuses on the dose applied in a specific use scenario, taking into consideration factors such as the amount of the substance used, the frequency and duration of use, the type of usage (leave-on or rinse-off), and the surface area covered. A deterministic approach can be used to address external exposure, as outlined in the SCCS Notes of Guidance (SCCS, 2023). It is also possible to refine exposure assessment by considering consumer habits and practices, particularly for combined and/or aggregate exposure scenarios (Tozer et al., 2019).

Depending on the type of safety assessment being conducted, the margin of safety (MoS) can be calculated using external exposure values. For NGRA, a margin of internal exposure (MoIE) or bioactivity:exposure ratio (BER) can be derived using data from NAMs. Therefore, external exposure values are translated into internal concentrations using PBK modelling. The uncertainty associated with the PBK-modelled internal concentration is usually estimated through sensitivity analysis and compared to measured benchmark values, whenever available. The input parameters required for PBK modelling can be derived from *in silico* and/or *in vitro* assays (Moxon et al., 2020). Fragki *et al.* (2023) applied a PBK model-facilitated reverse dosimetry approach for the translation of *in vitro* PFAS concentrations to external doses, allowing the translation of an *in vitro* point of departure (PoD) to an external dose that was compared with the actual dietary exposure in the population. It should be noted that this approach is deterministic and, therefore, operates with fixed parameters and does not explicitly account for uncertainty. Quantitative *in vitro* to *in vivo* extrapolation (QIVIVE) via reverse dosimetry is usually based on the plasma peak concentration (C_{max}) or plasma concentration at steady state (C_{ss}), and relies less frequently on the area under curve (AUC). However, because human exposure to PFAS is chronic and their elimination half-lives are rather long (*i.e.*, years), Fragki *et al.* (2023) relied on the AUC for a single 24-h *in vitro* exposure, extrapolated to a lifetime chronic (*i.e.*, 50-year) human oral exposure.

Working with realistic exposure estimates is essential for understanding the concentrations/exposures where bioactivity (or, where it can be delineated, adversity) occur and comparing them with the intended exposure in each use scenario. Considering the shift in paradigm from traditional animal-based studies to the use of NAMs in NGRA, the PoD used in the assessment can be based on bioactivity, such as changes at the molecular, cellular, or tissue levels that may not directly indicate adverse effects in humans, or on KEs and biomarkers that are clearly associated with adverse outcomes in humans. The MoIE or BER represents the ratio of the internal exposure to the selected PoD, it provides a measure of the potential risk associated with the exposure, and it is used to assess the safety of the cosmetic ingredient.

3.2. Exposure refinement

Refining exposure is another important aspect of ensuring human relevance in toxicological risk assessments. One approach to achieve this is through *in vitro* biokinetics, which can define effective concentrations rather than relying on nominal concentrations for bioactivity assays.

Several published case studies adhere to this principle, which emphasizes the identification of exposures relevant to the specific use context. In a case-study of coumarin (Baltazar et al., 2020), a risk-based and exposure-led approach was taken, incorporating a robust PBK model that considered population variability. The assessment included a conservative estimate of consumer exposure based on a worst-case scenario, and a refined PBK model was developed using *in vitro* absorption, distribution, metabolism, and excretion (ADME) parameters. The systemic exposure was expressed as the distribution of plasma peak concentration (C_{max}) across a population, accounting for inter-individual variability. The PBK predictions were validated using clinical data.

Similarly, for methyldibromo glutaronitrile (MDBGN) and propylparaben assessed for skin sensitization (Gilmour et al., 2020; Vandecasteele et al., 2021), the case studies defined a clear use scenario, and an exposure-led assessment was conducted using a deterministic dose estimate.

In a case study of caffeine (Bury et al., 2021; OECD, 2020b), exposure from both dermal and oral routes was assessed using a deterministic approach. The estimated dose was then translated into an internal concentration using PBK modelling, which incorporated *in vitro* data and accounted for population variability. The robustness of the predictions was confirmed with clinical data.

Furthermore, in the assessment of the systemic effects of propylparaben (OECD, 2020a; Ouédraogo et al., 2022), both deterministic and probabilistic aggregate exposures were considered for the external dose. PBK modelling, validated with clinical data, was used to characterise human exposure.

In a case study on phenoxyethanol (Hewitt et al., 2022; OECD, 2021b), the exposure characterization revealed the presence of the main metabolite (*i.e.*, phenoxyacetic acid) in the kidney. However, further investigation of bioactivity in this organ was not conducted, representing a limitation in that study. PBK modelling with *in vitro* ADME parameters, accounting for population variability, was employed together with comprehensive metabolism characterization. *In vitro* biokinetics were utilized to identify the types and extent of species present in the toxicodynamics assays.

Case studies on coumarin (Baltazar et al., 2020; Reynolds et al., 2021) and phenoxyethanol (Hewitt et al., 2022) evaluating systemic toxicity considered a worst-case scenario for the applied dose. A tiered approach was employed for PBK modelling, starting with *in silico* input parameters and refining the model with *in vitro* hepatic clearance measurements, including specific enzyme inhibitors. This approach provided evidence regarding the main clearance mechanism of coumarin, and population variability was accounted for in the estimation of internal concentrations. In the assessment of skin sensitization, two scenarios (*i.e.*, face cream and deodorant) highlighted the significance of exposure assessment, which could result in varying levels of risk.

Exposure refinement and justification of the metric relevant for the risk assessment were also demonstrated in a case study on PFAS (Fragki et al., 2023), where a quantitative *in vitro* to *in vivo* extrapolation (QIVIVE) was performed. Three *in vitro* dose metrics (the concentration in the medium, the medium concentration adjusted by the liver:blood partition coefficient, and the intracellular concentration) were each measured and applied as potential surrogates for a PBK- modelled hepatic concentration used in the reverse dosimetry PBK modelling to predict the corresponding oral human equivalent values. After comparing the results using the three different *in vitro* dose metrics, the authors provided a rationale of recommending the use of intracellular concentration in QIVIVE. However, the nominal concentrations in the medium corrected using the liver:blood partition coefficient were determined to be an acceptable alternative.

By considering external/internal exposures and population variability, as well as incorporating real-world data, utilizing NAMs in safety assessments and validating with clinical data can contribute a more accurate and representative understanding of human exposure, and better inform safety decision-making.

4. The risk assessment is hypothesis-driven

The hypothesis that underpins a NGRA may be highly specific (*e.g.*, related to whether interaction with a specific cellular target or biological pathway would result in an adverse health effect). Conversely, where there is no defined cellular target or biological pathway, the hypothesis may be more general. This concept was outlined in the US EPA's Next Generation Blueprint of Computational Toxicology (Thomas et al., 2019), which describes a tiered approach for hazard characterization. The first tier uses a broad-coverage screen across multiple cell types to evaluate bioactivity and to group chemicals based on similarity in potential hazards. Where no defined biological target/pathway is identified, a quantitative PoD for safety decision-making is estimated based on the absence of biological activity in the assays used. However, where the first tier does identify a specific biological target or pathway, this can be followed up with targeted assays and using knowledge (where available) from adverse outcome pathways. In this scenario, the points of departure may be based on the mode of action or responses in more advanced cell/tissue models.

For safety assessment purposes, it is crucial that the hypothesis is linked to an exposure level or use scenario of the chemical being assessed, as in a traditional risk assessment, because the objective of an NGRA is to make an exposure-led decision, not simply to identify bioactivity.

There are examples of NGRA case studies in the literature that illustrate both these scenarios. For example, systemic toxicity case studies on the use of both coumarin and phenoxyethanol in cosmetics (Baltazar et al., 2020; OECD, 2021b) have explored the concept of making a safety decision in the absence of a specific mode of action hypothesis. The hypothesis underpinning this type of NGRA is that if there is no bioactivity observed at consumer-relevant concentrations, there can be no adverse health effects (Baltazar et al., 2020). In the case study on phenoxyethanol, the overarching hypothesis was that systemic exposure to phenoxyethanol present at 1 % in body lotion will not cause any adverse health effects in consumers. Following the evaluation of available data, the authors concluded a lack of compelling evidence that phenoxyethanol was active in humans by any specific modes of action and, therefore, was considered to fit into the 'no defined biological target or pathway' category as described in the US EPA's blueprint for computational toxicology (Thomas et al., 2019).

Other IATA case studies have been developed that target different hypotheses. For example, the overarching hypothesis in a case study on caffeine was that caffeine and its analogues have a similar mode of action across species, but with different potencies, meaning that *in vivo* data from the analogues may be used to fill data gaps for caffeine (OECD, 2020b). This hypothesis was therefore centred on specific cellular events (antagonism of adenosine receptors, phosphodiesterase inhibition, modulation of GABA receptor action, and regulation of intracellular calcium levels). A challenge with this kind of approach is ensuring that the safety assessment is also protective of any biological effects of the target chemical (in this case caffeine) that are not exhibited by the source chemical(s). It is therefore important to ensure that the NAMs selected to determine the similarity of the mode of action are not only focussed on or biased by existing knowledge, but that they provide a broad enough screen to account for alternative modes of action.

In published NAM-based skin allergy case studies (Gilmour et al., 2020; Reynolds et al., 2021; Vandecasteele et al., 2021) scientific knowledge of the KEs leading to the induction of skin sensitization has been used to formulate specific mode of action hypotheses. In some

cases, this included consideration of metabolism of the parent compound to reduce uncertainty in the safety assessment. For example, in a case study on propylparaben (Vandecasteele et al., 2021), the biotransformation of the parent compound to 4-hydroxybenzoic acid was explored. This metabolite was predicted by the NAM and defined approach data to be non-sensitizing. There was a high concordance between the results of NAM studies on 4-hydroxybenzoic acid, whereas the NAM data for propylparaben itself was less concordant. Under a conservative WoE approach, the overall conclusion was that propylparaben was, at worst, a weak sensitizer.

In summary, different safety assessments may have very different hypotheses. It is therefore important for safety assessors to clearly articulate the hypothesis, to enable a reader/reviewer to understand whether the approaches used are appropriate to test it. One example of best practice in this area is the IATA case study on the read-across for dermal exposure to propylparaben in cosmetics (OECD, 2020a). This case study clearly articulated a read-across hypothesis, which was: *"The starting hypothesis for this category is that, based on their highly similar chemical structure, the target chemical propylparaben will have similar bioavailability and bioactivity as the source chemicals methyl, ethyl and butyl paraben."* For use in an exposure-led safety assessment, this would be considered a sub-hypothesis within the overarching safety assessment hypothesis relating to a use level and product type. For example:

Overarching safety assessment hypothesis

The use of x % of propylparaben in leave-on or rinse-off cosmetics does not induce skin sensitization.

Sub-hypothesis

Based on their highly similar chemical structure, the target chemical propylparaben has similar bioavailability and bioactivity as the source chemicals methylparaben, ethylparaben and butylparaben.

In this IATA, the read-across hypothesis was followed by a step-by-step breakdown of the considerations that underpin the hypothesis (or further sub-hypotheses), including chemical structure, available *in vivo* data, quantitative and qualitative metabolism, and potency trend:

Sub-hypothesis

The similar chemical structure and physicochemical characteristics result in similar bioavailability, metabolism and reactivity, which then result in similar biological and functional effects.

Sub-hypothesis

The available *in vivo* systemic toxicity data generally demonstrate the similar biological activity across the category. The same or similar mode of action across category members is responsible for the observed effects.

Sub-hypothesis

The parent category members are metabolized by ester hydrolysis via endogenous esterases in the skin or systemically after absorption, with all four parabens producing a common primary metabolite, 4-hydroxybenzoic acid (pHBA), and similar corresponding short linear chain alcohols.

Sub-hypothesis

The rate and extent of ester hydrolysis is similar across parabens, resulting in similar exposures to the common metabolite pHBA, which does not contribute significantly to the observed toxicity.

Sub-hypothesis

Chain length differences in the parent esters and in the other primary metabolite (*i.e.*, the linear aliphatic alcohol) across the parabens result in a predictable potency trend in observed effects across category members with increasing alkyl chain length.

Although this may appear complicated at first glance, this sub-delineation provides clarity of the overall structure of the safety assessment and the lines of evidence needed to make a conclusion, supporting the entire case study and demonstrating the suitability of the NAMs selected to test the hypothesis. Structuring the hypothesis in this way will greatly increase people's understanding of, and therefore the acceptance of, NGRA approaches.

5. The assessment is designed to prevent harm

When cosmetic ingredient safety assessments are called to move away from the traditional animal testing-based paradigms, questions have arisen around how reliable NAMs-based predictions are, and if the results from non-animal methods would provide adequate health protection. In fact, the premise of "protection, not prediction" sets the goal of many NGRAs, which is to provide the basis for decision-making to prevent harm to consumers at relevant exposures, instead of predicting toxicity through animal testing, which may not be necessary. NGRA takes the advantage of the development of various AOPs which indicate that some early cellular-level or sub-organoid biological activities (*i.e.*, KEs) precede the occurrence of adverse outcomes. Based on the concept that where no biological activity is predicted to occur at human-relevant exposures there can be no adversity (Dent et al., 2018; 2021), NGRA starts from the mechanism(s) associated with specific toxicological endpoints and relies on hypothesis-based *in vitro, in silico* and *in chemico* information to determine whether any cellular- or organ-level responses are likely to be induced at a human exposure-relevant concentration of a chemical. Recent NGRA case studies have offered some practical directions for implementing this principle.

5.1. Bioactivity characterization

Risk assessors should apply other relevant principles outlined in Dent et al. (2018) to evaluate bioactivity detected. For instance, a NAM-based PoD should be derived from robust studies that detect human-relevant bioactivities or signals.

Because mechanism-based testing can provide great sensitivity in detecting biological activities, NGRA supported by relevant, robust testing and (where available and appropriate) well-defined AOPs should provide for adequate confidence to prevent harm when there is no biological activity at concentrations relevant for the final consumer. Special consideration may be needed when biological activities relevant to adverse effects of concern (*e.g.*, cancer or developmental effects) are detected and may require further evaluation. In whole-animal studies, an adverse effect is defined as a "*test-item related change in the morphology, physiology, growth, development, reproduction or life span of the animal model that likely results in an impairment of functional capacity to maintain homeostasis and/or an impairment of the capacity to respond to an additional challenge*" (Palazzi et al., 2016). In contrast, there is no clear definition of adversity for *in vitro* or *ex vivo* tests leading to systemic toxicity. As demonstrated by the case study on coumarin (Baltazar et al., 2020), in which the PoD was based on biological activities rather than adverse effects, a safety determination was straightforward when the PoD_{NAM} provides adequate margin towards even a worst-case estimate of internal concentration (plasma C_{max}). However, when biological activities are detected at human-relevant exposure levels, methods to differentiate between an adaptive and an adverse response are needed to avoid over-conservative decision-making (Dent et al., 2021). Due to the complexity of biological responses to chemical stimulation, the evaluation would require the consideration of information from different types of mechanistic studies covering multiple signaling pathways and various stages of adverse effect development. As an example, a recently proposed framework of NAM-based developmental and reproductive toxicity (DART) outcome was applied to a case study of NGRA for caffeine (Rajagopal et al., 2022). In this study, *in vitro* biological activity

characterization was performed by collecting and integrating data from studies using a cell stress panel, high-throughput transcriptomics, and *in vitro* pharmacological profiling of broad receptor binding activities. These NAMs indicated that caffeine can antagonize the adenosine 2A receptor at internal concentrations that could be experienced through the consumption of caffeinated beverages (a known pharmacological action of caffeine). To prevent harm, it is important to assess whether this bioactivity could lead to an adverse health effect in consumers. Because antagonism of adenosine receptors could potentially link to concerns for cardiac and neuronal embryonic tissue development, the authors performed follow-up studies using induced pluripotent stems cells to examine the effect of caffeine on cell differentiation to cardiomyocytes and neuronal rosettes. As these studies were negative at consumer-relevant concentrations, they provided reassurance that the bioactivity observed at the level of the receptor would not be sufficient to cause an adverse effect in the windows of development covered by the assays.

5.2. Integration of information

Integration of information is essential to NGRA because the adverse effects (local or systemic) often involve multiple distinct biological events and pathways. It is unlikely that any single NAM testing could be adequate to support a reliable safety evaluation. The framework brought out by Gilmour et al. (2020) includes an initial tier involving a thorough review of the existing information and a WoE approach in the next steps to appraise the totality of information towards the potential of causing skin sensitization. Under this framework, all existing information, including *in silico, in chemico, in vitro, in vivo* and historical human clinical data should all be collected to make a WoE-based safety determination. If it is determined that the integrated information is still inadequate to conduct a risk assessment, further non-animal systems such as organ chips or organoid-based *in vitro* models may be conducted to fill the necessary data gaps. Using a case study of methyldibromo glutaronitrile (MDBGN), the authors demonstrated the practice of integrating a broad range of NAM evidence with a good coverage of different sequential KEs in the skin sensitization AOP, as well as using defined approaches (DAs) to derive a PoD. A similar tiered framework was also applied in an IATA case study of NGRA of phenoxyethanol in systemic toxicity assessment (OECD, 2021b). Vandecasteele et al. (2021) presented a case study of propylparaben in a situation where the negative skin sensitization prediction by sequential stacking tier testing DA[2] was accompanied with some positive NAM input data. Although a conservative, *ab initio* NGRA (As an exercise that intentionally did not use any pre-existing data, but was performed on a data-rich compound with a well-established safety outcome which allowed the pre-existing data to be used as a reference to determine the relevance and the reliability of the prediction) predicted that propylparaben was at worst a weak sensitizer, the authors demonstrated the use of additional NAM, DA, animal, and human data for structurally related parabens to address the uncertainty introduced by the discordant *in vitro* NAM results. The read-across data supported the non-sensitizer DA prediction, thus providing further confidence in the WoE that 0.2 % propylparaben in face cream is a low risk for inducing skin sensitization.

Additionally, whether a response or biological signal will ultimately lead to adverse events also depends on concentration-response relationship and the quantitative relationship between KEs. While still a challenging area under development, efforts have been made to establish quantitative AOP models (qAOPs) relevant for decision-making (Perkins et al., 2019). Aiming to link exposure to the amount of chemical needed to cause a PoD in an AOP, qAOPs provide quantitative descriptors for key event relationships (KERs) by analyzing concentration-response data using statistical, Bayesian, regressional or other methods. Combined with PBPK modelling, qAOP models have the potential to estimate an *in vivo* PoD that can be used to evaluate the safety of certain external exposures. Exercises on qAOPs, with steps of data integration and modelling have been conducted for skin sensitization, neurotoxicity and carcinogenicity (Paini et al., 2022). However, qAOPs rely on much more information on the quantitative relationships between KEs than is typically available (Sinitsyn et al., 2022). At the same time, the validation and harmonization of qAOP models need further research (Spinu et al., 2020). Thus, in the spirit of NGRA as a tiered and iterative approach, it is likely that this kind of resource-intensive approach would be reserved for the higher tiers of assessment, if there is a need to distinguish between biological activity and adversity on a quantitative basis. The context of the assessment (set out clearly during problem formulation) therefore dictates how the output can be used. For example, where the safety assessment relies on demonstrating a lack of any significant bioactivity at human-relevant exposure concentrations (*e.g.* Baltazar et al., 2020; OECD 2021b), this does not inform hazard identification/characterization, and should therefore not be used to take hazard-based regulatory action.

5.3. Being protective

With the goal of preventing harm, NGRA generally applies conservative parameters in exposure estimates to being protective, such as for high users and vulnerable subpopulations. This is commonly seen in almost all published NGRA case studies. For example, in the case study on phenoxyethanol conducted by OECD (2021b), the exposure considered was dermal application to consumers when 1 % phenoxyethanol is included in a body lotion. For the applied dose, exposure assessment worst-case (conservative) assumptions were made regarding product use. Consumer use and physiological data for females were used to ensure the safety assessment was conservative, because they have lower mean body weights and greater mean use levels than males. The exposure scenario assumes all the applied body lotion was left in contact with skin until the next application. In the absence of any skin penetration data identified, the dermal absorption was assumed to be 100 %. The approach was intended to be protective of human health rather than predictive of apical endpoints.

5.4. Iterative approach

The principle of using a tiered and iterative approach for NGRA, as outlined in Dent et al. (2018) also fits into its overall goal of preventing harm. A tiered approach starts with information that is the least resource-intensive (typically the most conservative) parameters. Based on the bioactivity:exposure ratio (BER) or margin of safety (MoS), the risk assessment can either stop because an adequate level of human health protection is achieved, or a more refined, second-tier assessment needs to be conducted. The criteria (*e.g.*, BER cut-offs) that would trigger refinement of the assessment have not been established for all use cases, and therefore need to be determined based on the specific context of the assessment, including the NAMs used. For example, for one possible prototype NAM toolbox and decision model for systemic toxicity, threshold BERs ranging between 110 and 2.5 have been proposed, with the choice of threshold dependent on the level of confidence in the exposure estimate (Middleton et al., 2022). Evaluation of this approach has found it is able to correctly identify the majority (>90 %) of high-risk exposure scenarios when tested with an additional set of 38 chemicals, providing a similar level of protection to *in vivo* animal studies (Cable et al., 2025). Such analyses are critical to ensure the

[2] "Sequential stacking tier testing defined approach" refers to a systematic, multi-tiered and data-driven method used for assessing the potential of a substance to cause skin sensitization, with a focus on high confidence in the predictive outcomes. This approach integrates various types of data and testing methods in a sequential manner. Each tier of assessment adds a layer of information, and the results from one tier can determine if further testing in the next tier is necessary.

selected NAM toolbox can correctly identify where a low-risk conclusion cannot be reached and either further refinement or lowering of use levels is needed. In practice, where a risk assessment needs to be refined, an important area to focus on is the exposure estimate. SCCS (SCCS, 2023) provides guidance and examples of conducting a refined (such as probabilistic) exposure assessment, and tiered approaches have already been defined for improving confidence in internal exposure (PBK) estimates to dermally-applied cosmetic ingredients (Moxon et al., 2020). With further precision derived from the refined assessment, more confidence is gained in making a safety decision.

Overall, instead of finding the dose level (*e.g.*, NOAEL) administered to the animals that does not result in adverse effects (*i.e.*, apical endpoints), NGRA relies on the mechanistic knowledge of toxicity and integrates NAM data from different sources to achieve the goal of preventing harm. This requires many different robust and relevant NAMs, some to be developed or refined, to provide sufficient biological coverage to ensure that the NGRA is at least as protective as the traditional (animal-based) approach.

6. Discussion and conclusion

As multiple organizations and regulatory agencies published strategies or roadmaps to help advance the acceptance and use of NAMs in making regulatory decisions (van der Zalm et al., 2022), an increasing number of NGRA case studies are appearing in the literature. The NGRA case studies described here do not represent an exhaustive list. However, it is clear that those claiming to follow the ICCR Principles of NGRA were transparent and well conducted.

This document highlights some of the best practices, providing a useful resource to cosmetic safety assessors striving to ensure their risk assessments are human-relevant, exposure-led, hypothesis-driven, and designed to prevent harm. Even in some of the best-conducted assessments, how these four goals of NGRA were achieved was not always explicitly described. The ICCR therefore encourages all those performing NGRA to ensure the assessment explicitly meets the ICCR Principles of NGRA. How this has been achieved needs to be explained transparently and communicated clearly in the risk assessment documentation.

One important observation from the literature is that many NGRA case studies do not aim to predict specific hazards associated with substance exposure, but rather are designed to ensure that at relevant exposures there will be no meaningful bioactivity (*e.g.*, Baltazar et al., 2020; OECD 2021b). However, other case studies tended to be more predictive of specific adverse effects such as skin sensitization (Gilmour et al., 2020) or hepatotoxicity (Fragki et al., 2023). The type of safety assessment that has been performed ('protective' or 'predictive') as well as its robustness and level of uncertainty set the context of its regulatory use, including screening, prioritization and so on. For example, while a 'protective' approach may be useful for assuring the safety of specific exposures based on the lack of significant bioactivity, it would be inappropriate to use it as a basis for regulatory restriction such as an ingredient ban, simply because in a 'protective' NGRA no specific hazards are identified. On the other hand, while predictive data on some fragrance allergens resulted in the 2023 amendment of the EU cosmetic regulation on the labelling of these ingredients, not all potential hazards 'predicted' by NAMs would warrant regulatory restriction or ban. Nevertheless, the goal of both types of assessment is to facilitate risk managers and regulators to make informed decisions. It is clear that, in addition to the general best practice described in this document which aims to encourage safety assessments to explicitly reflect the ICCR principles, there is a need for more specific NGRA guidance. For example, of the two good quality case studies on propylparaben, one covered skin sensitization (Vandecasteele et al., 2021) and one covered systemic toxicity/endocrine activity (OECD 2020a). However, to date no guidance exists on how a safety assessor would be confident that their safety assessment ensures that all relevant biological activities and uncertainties are covered. This could leave safety assessors and reviewers uncertain as to the overall validity of the safety assessment conclusion, and undermine confidence in NGRA. It is therefore important to develop this next level of guidance, moving even further from principles into application. This could begin with the development of more specific guidance and workflows covering the health effects of concern for cosmetic ingredients, starting with local effects (such as skin and eye irritation, skin sensitization and local effects in the lung) where many different NAMs and in some cases DAs exist, and moving on to the more complex health effects such as systemic, developmental and reproductive toxicity as more consensus develops in these areas. Such specific guidance could provide the basis for constructive dialogue between data submitters and regulatory authorities to reach consensus on how confidence can be built in the use of NAMs for regulatory decision-making. This work needs to be considered a priority to ensure that the benefits of NAMs are fully realized, and the move to new safety assessment paradigms is made in a transparent and responsible manner.

Disclaimer

The content of this publication does not necessarily reflect the views or policies of the Brazilian Health Regulatory Agency (ANVISA), Health Canada (HC), the Taiwan Food and Drug Administration (TFDA), the Japanese Ministry of Health, Labour and Welfare (MHLW), the US Environmental Protection Agency (US EPA), the US Food and Drug Administration (US FDA) Department of Health and Human Services, or the European Commission (EC), nor does mention of commercial products, or organizations imply endorsement by the regulatory authorities of Brazil, Canada, Chinese Taipei, Japan and United States or the European Commission. This paper reflects the current thinking and experience of the authors. As with any ICCR work product, any best practice does not imply adoption/implementation within individual ICCR jurisdictions.

CRediT authorship contribution statement

Eric Vaillancourt: Writing – review & editing, Writing – original draft, Project administration, Conceptualization. **Alaa Afdal Mohamed:** Writing – review & editing, Writing – original draft. **Jay Ansell:** Writing – review & editing, Writing – original draft, Conceptualization. **Takao Ashikaga:** Writing – review & editing, Writing – original draft, Conceptualization. **Ahmed Ayad:** Writing – review & editing. **Hadeel Ayman:** Writing – review & editing, Writing – original draft. **Marize Valadares Campos:** Writing – review & editing, Writing – original draft, Conceptualization. **Han-Sheng Chien:** Writing – review & editing, Writing – original draft, Conceptualization. **Arianna Giusti:** Writing – review & editing, Writing – original draft, Conceptualization. **Masato Hatao:** Writing – review & editing, Writing – original draft, Conceptualization. **Frances Hill:** Writing – original draft, Writing – review & editing, Conceptualization. **Abeer Khattab:** Writing – review & editing. **Reinhard Kreiling:** Writing – review & editing, Conceptualization. **Ah Young Lee:** Writing – review & editing. **Miao Li:** Writing – review & editing, Writing – original draft, Conceptualization. **Pinpin Lin:** Writing – review & editing, Writing – original draft, Conceptualization. **Yen-Chun Lin:** Writing – review & editing, Writing – original draft, Conceptualization. **Victor Mencarelli:** Writing – review & editing, Writing – original draft, Conceptualization. **Stanley Milstein:** Writing – review & editing, Conceptualization. **Beta P. Montemayor:** Writing – review & editing, Writing – original draft, Conceptualization. **Carolina Motter Catarino:** Writing – review & editing, Writing – original draft, Conceptualization. **Rawda Negm:** Writing – review & editing. **Gladys Ouédraogo:** Writing – review & editing, Writing – original draft, Conceptualization. **Eun Jae Park:** Writing – review & editing, Writing – original draft, Conceptualization. **Mustafa Varçin:** Writing – review & editing, Writing – original draft, Conceptualization. **Tzu-Yi Yang:** Writing – review & editing, Writing – original draft, Conceptualization. **Yu Janet Zang:** Writing – review & editing, Writing – original draft,

Conceptualization. **Matthew Philip Dent:** Writing – review & editing, Writing – original draft, Project administration, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marize Valadares Campos, Gladys Ouédraogo and Matthew P. Dent serve on the Editorial Board of NAM Journal. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our gratitude to previous members of the International Cooperation on Cosmetics Regulation (ICCR)'s Integrated Strategies for Safety Assessment II Regulator-Industry Joint Working Group as well as to the ICCR Steering Committee members who have contributed insightful discussions and comments on the manuscript draft.

Data availability

No data was used for the research described in the article.

References

Baltazar, M.T., Cable, S., Carmichael, P.L., Cubberley, R., Cull, T., Delagrange, M., Dent, M.P., Hatherell, S., Houghton, J., Kukic, P., Li, H., Lee, M.Y., Malcomber, S., Middleton, A.M., Moxon, T.E., Nathanail, A.V., Nicol, B., Pendlington, R., Reynolds, G., Reynolds, J., White, A., Westmoreland, C., 2020. A next-generation risk assessment case study for Coumarin in cosmetic products. Toxicol. Sci. 176 (1), 236–252. https://doi.org/10.1093/toxsci/kfaa048.

Bartsch, R., Brinkmann, B., Jahnke, G., Laube, B., Lohmann, R., Michaelsen, S., Neumann, I., 2018. Human relevance of follicular thyroid tumors in rodents caused by non-genotoxic substances. Regul. Toxicol. Pharmacol. 98, 199–208. https://doi.org/10.1016/j.yrtph.2018.07.025.

Berggren, E., White, A., Ouédraogo, G., Paini, A., Richarz, A.N., Bois, F.Y., Exner, T., Leite, S., van Grunsven, L.A., Worth, A., Mahony, C., 2017. Ab initio chemical safety assessment: a workflow based on exposure considerations and non-animal methods. Comp. Toxicol. 4, 31–44. https://doi.org/10.1016/j.comtox.2017.10.001.

Bury, D., Alexander-White, C., Clewell, HJ 3rd, Cronin, M., Desprez, B., Detroyer, A., Efremenko, A., Firman, J., Hack, E., Hewitt, N.J., Kenna, G., Klaric, M., Lester, C., Mahony, C., Ouédraogo, G., Paini, A., Schepky, A., 2021. New framework for a non-animal approach adequately assures the safety of cosmetic ingredients - a case study on caffeine. Regul. Toxicol. Pharmacol. 123, 104931. https://doi.org/10.1016/j.yrtph.2021.104931.

Cable, S., Baltazar, M.T., Bunglawala, F., Carmichael, P.L., Contreas, L., Dent, M.P., Houghton, J., Kukic, P., Malcomber, S., Nicol, B., Przybylak, K.R., Punt, A., Reynolds, G., Reynolds, J., Scott, S., Tang, D., Middleton, A.M., 2025. Advancing systemic toxicity risk assessment: evaluation of a NAM-based toolbox approach. Toxicol. Sci. 204 (1), 79–95. https://doi.org/10.1093/toxsci/kfae159.

Chamanza, R., Wright, J.A., 2015. A review of the comparative anatomy, histology, physiology and pathology of the nasal cavity of rats, mice, dogs and non-human primates. Relevance to inhalation toxicology and human health risk assessment. J. Comp. Pathol. 153 (4), 287–314. https://doi.org/10.1016/j.jcpa.2015.08.009.

Cook, J.C., Klinefelter, G.R., Hardisty, J.F., Sharpe, R.M., Foster, P.M.D., 1999. Rodent Leydig cell tumorigenesis: a review of the physiology, pathology, mechanisms, and relevance to humans. Crit. Rev. Toxicol. 29 (2), 169–261. https://doi.org/10.1080/10408449991349203.

Cunha, G.R., Sinclair, A., Risbridger, G., Hutson, J., Baskin, L.S., 2015. Current understanding of hypospadias: relevance of animal models. Nat. Rev. Urol. 12, 271–280. https://doi.org/10.1038/nrurol.2015.57.

Dent, M., Teixeira, Amaral R, Amores, Da Silva P, Ansell, J., Boisleve, F., Hatao, M., Hirose, A., Kasai, Y., Kern, P., Kreiling, R., Milstein, S., Montemayor, B., Oliveira, J., Richarz, A., Taalman, R., Vaillancourt, E., Verma, R., Vieira, O'Reilly Cabral Posada N, Weiss, C., Kojima, H., 2018. Principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients. Comp. Toxicol. 7, 20–26. https://doi.org/10.1016/j.comtox.2018.06.001.

Dent, M.P., Vaillancourt, E., Thomas, R.S., Carmichael, P.L., Ouédraogo, G., Kojima, H., Barroso, J., Ansell, J., Barton-Maclaren, T.S., Bennekou, S.H., Boekelheide, K., Ezendam, J., Field, J., Fitzpatrick, S., Hatao, M., Kreiling, R., Lorencini, M., Mahony, C., Montemayor, B., Mazaro-Costa, R., Oliveira, J., Rogiers, V., Smegal, D., Taalman, R., Tokura, Y., Verma, R., Willett, C., Yang, C., 2021. Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients. *Regul. Toxicol. Pharmacol* 125, 105026. https://doi.org/10.1016/j.yrtph.2021.105026.

Desalegn, A., Bopp, S., Asturiol, D., Lamon, L., Worth, A., Paini, A., 2019. Role of physiologically based kinetic modelling in addressing environmental chemical mixtures – a review. Comp. Toxicol. 10, 158–168. https://doi.org/10.1016/j.comtox.2018.09.001.

Ebmeyer, J., Najjar, A., Lange, D., Boettcher, M., Voß, S., Brandmair, K., Meinhardt, J., Kuehnl, J., Hewitt, N.J., Krueger, C.T., Schepky, A., 2024. Next generation risk assessment: an *ab initio* case study to assess the systemic safety of the cosmetic ingredient, benzyl salicylate, after dermal exposure. Front. Pharmacol. 15, 1345992. https://doi.org/10.3389/fphar.2024.1345992.

EU Cosmetics Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products (recast) (Text with EEA relevance). Offic. J. Euro. Union, 2009. https://eur-lex.europa.eu/eli/reg/2009/1223/oj.

Fragki, S., Louisse, J., Bokkers, B., Luijten, M., Peijnenburg, A., Rijkers, D., Piersma, A.H., Zeilmaker, M.J., 2023. New approach methodologies: a quantitative *in vitro* to *in vivo* extrapolation case study with PFASs. Food Chem. Toxicol. 172, 113559. https://doi.org/10.1016/j.fct.2022.113559.

Gilmour, N., Kern, P.S., Alépée, N., Boislève, F., Bury, D., Clouet, E., Hirota, M., Hoffmann, S., Kühnl, J., Lalko, J.F., Mewes, K., Miyazawa, M., Nishida, H., Osmani, A., Petersohn, D., Sekine, S., van Vliet, E., Klaric, M., 2020. Development of a next generation risk assessment framework for the evaluation of skin sensitisation of cosmetic ingredients. *Regul. Toxicol. Pharmacol* 116, 104721. https://doi.org/10.1016/j.yrtph.2020.104721.

Hall, B., Tozer, S., Safford, B., Coroama, M., Steiling, W., Leneveu-Duchemin, M.C., McNamara, C., Gibney, M., 2007. European consumer exposure to cosmetic products, a framework for conducting population exposure assessments. Food Chem. Toxicol. 45 (11), 2097–2108. https://doi.org/10.1016/j.fct.2007.06.017.

Hatherell, S., Baltazar, M.T., Reynolds, J., Carmichael, P.L., Dent, M., Li, H., Ryder, S., White, A., Walker, P., Middleton, A.M., 2020. Identifying and characterizing stress pathways of concern for consumer safety in next-generation risk assessment. Toxicol. Sci. 176 (1), 11–33. https://doi.org/10.1093/toxsci/kfaa054.

Hewitt, N.J., Troutman, J., Przibilla, J., Schepky, A., Ouédraogo, G., Mahony, C., Kenna, G., Varçin, M., Dent, M.P., 2022. Use of *in vitro* metabolism and biokinetics assays to refine predicted *in vivo* and *in vitro* internal exposure to the cosmetic ingredient, phenoxyethanol, for use in risk assessment. Regul. Toxicol. Pharmacol. 131, 105132. https://doi.org/10.1016/j.yrtph.2022.105132.

Krewski, D., Jr, Acosta D, Andersen, M., Anderson, H., Bailar III, J.C., Boekelheide, K., Brent, R., Charnley, G., Cheung, V.G., Jr, Green S, Kelsey, K.T., Kerkvliet, N.I., Li, A. A., McCray, L., Meyer, O., Patterson, R.D., Pennie, W., Scala, R.A., Solomon, G.M., Stephens, M., Yager, J., Zeise, L., Staff of Committee on Toxicity Testing and Assessment of Environmental Agents, 2010. Toxicity testing in the 21st century: a vision and a strategy. J. Toxicol. Environ. Health B 13 (2–4), 51–138. https://doi.org/10.1080/10937404.2010.483176.

Leist, M., Hartung, T., 2013. Inflammatory findings on species extrapolations: humans are definitely no 70-kg mice. Arch. Toxicol. 87 (4), 563–567. https://doi.org/10.1007/s00204-013-1038-0.

Lichtenstein, D., Luckert, C., Alarcan, J., Sousa, G., Gioutlakis, M., Katsanou, E.S., Kanstantinidou, P., Machera, K., Milani, E.S., Peijnenburg, A., Rahmani, R., Rijkers, D., Spyropoulou, A., Stamou, M., Stoopen, G., Sturla, S.J., Wollscheid, B., Zucchini-Pascal, N., Braeuning, A., Lampen, A., 2020. An adverse outcome pathway-based approach to assess steatotic mixture effects of hepatotoxic pesticides *in vitro*. Food Chem. Toxicol. 139, 111283. https://doi.org/10.1016/j.fct.2020.111283.

Middleton, A.M., Reynolds, J., Cable, S., Baltazar, M.T., Li, H., Bevan, S., Carmichael, P. L., Dent, M.P., Hatherell, S., Houghton, J., Kukic, P., Lidell, M., Malcomber, S., Nicol, B., Park, B., Paten, H., Scott, S., Sparham, C., Walker, P., White, A., 2022. Are non-animal systemic safety assessments protective? A toolbox and workflow. *Toxicol. Sci* 189 (1), 124–147. https://doi.org/10.1093/toxsci/kfac068.

Moxon, T.E., Li, H., Lee, M.Y., Piechota, P., Nicol, B., Pickles, J., Pendlington, R., Sorrell, I., Baltazar, M.T., 2020. Application of physiologically based kinetic (PBK) modelling in the next generation risk assessment of dermally applied consumer products. *Toxicol. In Vitro* 63, 104746. https://doi.org/10.1016/j.tiv.2019.104746.

OECD, 2004. OECD test Guideline 428 for testing of chemicals (TG 428) - Skin absorption: *in vitro* method. https://doi.org/10.1787/9789264071087-en.

OECD, 2020a. Case study on use of an integrated approach to testing and assessment (IATA) and new approach methods to inform a theoretical read-across for dermal exposure to propylparaben from cosmetics. Ser. Test. Assess. No. 320. Available at https://one.oecd.org/document/ENV/JM/MONO(2020)16/en/pdf.

OECD, 2020b. Case study on the use of integrated Approaches for testing and assessment for systemic toxicity arising from cosmetic exposure to caffeine. Ser. Test. Assessm. (321) Available at https://one.oecd.org/document/ENV/JM/MONO(2020)17/en/pdf.

OECD, 2021a. OECD test Guideline 497 for testing of chemicals (TG 497) –Defined approaches for skin sensitisation. DOI: https://doi.org/10.1787/b92879a4-en.

OECD, 2021b. Case study on use of an Integrated approach for testing and assessment (IATA) for systemic toxicity of phenoxyethanol when included at 1% in a body lotion. Ser. Test. Assessm. (349) Available at https://one.oecd.org/document/ENV/CBC/MONO(2021)35/En/pdf.

Ouédraogo, G., Alexander-White, C., Bury, D., Clewell, HJ 3rd, Cronin, M., Cull, T., Dent, M., Desprez, B., Detroyer, A., Ellison, C., Giammanco, S., Hack, E., Hewitt, N. J., Kenna, G., Klaric, M., Kreiling, R., Lester, C., Mahony, C., Mombelli, E., Naciff, J., O'Brien, J., Schepky, A., Tozer, S., van der Burg, B., van Vugt-Lussenburg, B., Stuard, S., 2022. Read-across and new approach methodologies applied in a 10-step framework for cosmetics safety assessment - a case study with parabens. *Regul. Toxicol. Pharmacol* 132, 105161. https://doi.org/10.1016/j.yrtph.2022.105161.

Paini, A., Campia, I., Cronin, M.T.D., Asturiol, D., Ceriani, L., Exner, T.E., Gao, W., Gomes, C., Kruisselbrink, J., Martens, M., Meek, M.E.B., Pamies, D., Pletz, J., Scholz, S., Schüttler, A., Spînu, N., Villeneuve, D.L., Wittwehr, C., Worth, A., Luijten, M., 2022. Towards a qAOP framework for predictive toxicology - linking data to decisions. Comp. Toxicol. 21, 100195. https://doi.org/10.1016/j.comtox.2021.100195.

Palazzi, X., Burkhardt, J.E., Caplain, H., Dellarco, V., Fant, P., Foster, J.R., Francke, S., Germann, P., Gröters, S., Harada, T., Harleman, J., Inui, K., Kaufmann, W., Lenz, B., Nagai, H., Pohlmeyer-Esch, G., Schulte, A., Skydsgaard, M., Tomlinson, L., Wood, C. E., Yoshida, M., 2016. Characterizing "adversity" of pathology findings in nonclinical toxicity studies: results from the 4th ESTP International Expert Workshop. Toxicol. Pathol. 44 (6), 810–824. https://doi.org/10.1177/0192623316642527.

Palmer, J.A., Smith, A.M., Egnash, L.A., Conard, K.R., West, P.R., Burrier, R.E., Donley, E. L.R., Kirchner, F.R., 2013. Establishment and assessment of a new human embryonic stem cell-based biomarker assay for developmental toxicity screening. Dev. Reprod. Toxicol. Birth Defects Res. B 98 (4), 343–363. https://doi.org/10.1002/bdrb.21078.

Paul-Friedman, K., Gagne, M., Loo, L-H, Karamertzanis, P., Netzeva, T., Sobanski, T., Franzosa, J.A., Richard, A.M., Lougee, R.R., Gissi, A., Lee, J.Y.J., Angrish, M., Dorne, J.L., Foster, S., Raffaele, K., Bahadori, T., Gwinn, M.R., Lambert, J., Whelan, M., Rasenberg, M., Barton-Maclaren, T., Thomas, R.S., 2020. Utility of *in vitro* bioactivity as a lower bound estimate of *in vivo* adverse effect levels and in risk-based prioritization. *Toxicol. Sci.* 173 (1), 202–225. https://doi.org/10.1093/toxsci/kfz201.

Perkins, E.J., Ashauer, R., Burgoon, L., Conolly, R., Landesmann, B., Mackay, C., Murphy, C.A., Pollesch, N., Wheeler, J.R., Zupanic, A., Scholz, S., 2019. Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environ. Toxicol. Chem* 38 (9), 1850–1865. https://doi.org/10.1002/etc.4505.

Rajagopal, R., Baltazar, M.T., Carmichael, P.L., Dent, M.P., Head, J., Li, H., Muller, I., Reynolds, J., Sadh, K., Simpson, W., Spriggs, S., White, A., Kukic, P., 2022. Beyond AOPs: a mechanistic evaluation of NAMs in DART testing. Front. Toxicol. 4, 838466. https://doi.org/10.3389/ftox.2022.838466.

Reynolds, G., Reynolds, J., Gilmour, N., Cubberley, R., Spriggs, S., Aptula, A., Przybylak, K., Windebank, S., Maxwell, G., Baltazar, M.T., 2021. A hypothetical skin sensitisation next generation risk assessment for coumarin in cosmetic products. *Regul. Toxicol. Pharmacol* 127, 105075. https://doi.org/10.1016/j.yrtph.2021.105075.

Rogiers, V., Benfenati, E., Bernauer, U., Bodin, L., Carmichael, P., Chaudhry, Q., Coenraads, P.J., Cronin, M.T.D., Dent, M., Dusinska, M., Ellison, C., Ezendam, J., Gaffet, E., Galli, C.L., Goebel, C., Granum, B., Hollnagel, H.M., Kern, P.S., Kosemund-Meynen, K., Ouédraogo, G., Panteri, E., Rousselle, C., Stepnik, M., Vanhaecke, T., von Goetz, N., Worth, A., 2020. The way forward for assessing the human health safety of cosmetics in the EU – workshop proceedings. Toxicology 436, 152421. https://doi.org/10.1016/j.tox.2020.152421.

Rovida, C., Hartung, T., 2009. Re-evaluation of animal numbers and costs for *in vivo* tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t^4). ALTEX 26 (3), 187–208. https://doi.org/10.14573/altex.2009.3.187.

Safford, R.J., 2008. The dermal sensitisation threshold—A TTC approach for allergic contact dermatitis. Regul. Toxicol. Pharmacol. 51 (2), 195–200. https://doi.org/10.1016/j.yrtph.2008.02.010.

SCCS, 2023. Notes of guidance for the testing of cosmetic ingredients and their safety evaluation - 12th revision. Available at: https://health.ec.europa.eu/system/files/2023-12/sccs_o_273_final.pdf.

Sepehri, S., De Win, D., Heymans, A., Van Goethem, F., Rodrigues, R.M., Rogiers, V., Vanhaecke, T., 2025. Next generation risk assessment of hair dye HC yellow no. 13: ensuring protection from liver steatogenic effects. *Regul. Toxicol. Pharmacol* 159, 105794. https://doi.org/10.1016/j.yrtph.2025.105794.

Sinitsyn, D., Garcia-Reyero, N., Wattanabe, K.H., 2022. From qualitative to quantitative AOP: a case study of neurodegeneration. Front. Toxicol. 4, 838729. https://doi.org/10.3389/ftox.2022.838729.

Spinu, N., Cronin, M.T.D., Enoch, S.J., Madden, J.C., Worth, A.P., 2020. Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. Arch. Toxicol. 94, 1497–1510. https://doi.org/10.1007/s00204-020-02774-7.

Thomas, R.S., Bahadori, T., Buckley, T.J., Cowden, J., Deisenroth, C., Dionisio, K.L., Frithsen, J.B., Grulke, C.M., Gwinn, M.R., Harill, J.A., Higuchi, M., Huock, K.A., Hughes, M.F., Hunter III, S.E., Isaacs, K.K., Judson, R.S., Knudsen, T.B., Lambert, J. C., Linnenbrink, M., Martin, T.M., Newton, S.R., Padilla, S., Patlewicz, G., Paul-Friedman, K., Phillips, K.A., Richard, A.M., Sams, R., Shafer, T.J., Setzer, R.W., Shah, I., Simmons, J.E., Simmons, S.O., Singh, A., Sobus, J.R., Strynar, M., Swank, A., Torneo-Valez, R., Ulrich, E.M., Villeneuve, D.L., Wambaugh, J.F., Wetmore, B.A., Williams, A., 2019. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. Toxicol. Sci. 163 (2), 317–322. https://doi.org/10.1093/toxsci/kfz058.

Tozer, S., O'Mahony, C., Hannah, J., O'Brien, J., Kelly, S., Kosemund-Meynen, K., Alexander-White, C., 2019. Aggregate exposure modelling of vitamin A from cosmetic products, diet and food supplements. *Food Chem. Toxicol* 131, 110549. https://doi.org/10.1016/j.fct.2019.05.057.

Vandecasteele, H.A., Gautier, F., Tourneix, F., Vliet, E.V., Bury, D., Alépée, N., 2021. Next generation risk assessment for skin sensitisation: a case study with propylparaben. Regul. Toxicol. Pharmacol. 123, 104936. https://doi.org/10.1016/j.yrtph.2021.104936.

van der Zalm, A.J., Barroso, J., Browne, P., Casey, W., Gordon, J., Henry, T.R., Kleinstreuer, N.C., Lowit, A.B., Perron, M., Clippinger, A.J., 2022. A framework for establishing scientific confidence in new approach methodologies. Arch. Toxicol. 96, 2865–2879. https://doi.org/10.1007/s00204-022-03365-4.

Wu, S., Fisher, J., Naciff, J., Laufersweiler, M., Lester, C., Daston, G., Blackburn, K., 2013. Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants. Chem. Res. Toxicol. 26 (12), 1840–1861. https://doi.org/10.1021/tx400226u.

Zurlinden, T.J., Saili, K.S., Rush, N., Kothiya, P., Judson, R.S., Houck, K.A., Hunter, E.S., Baker, N.C., Palmer, J.A., Thomas, R.S., Knudsen, T.B., 2020. Profiling the ToxCast library with a pluripotent human (H9) stem cell line-based biomarker assay for developmental toxicity. Toxicol. Sci. 174 (2), 189–209. https://doi.org/10.1093/toxsci/kfaa014.