Regulatory Toxicology and Pharmacology 112 (2020) 104592

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

An evaluation framework for new approach methodologies (NAMs) for human health safety assessment

Stanley T. Parish[a,*], Michael Aschner[b], Warren Casey[c], Marco Corvaro[d], Michelle R. Embry[a], Suzanne Fitzpatrick[e], Darren Kidd[f], Nicole C. Kleinstreuer[c], Beatriz Silva Lima[g], Raja S. Settivari[h], Douglas C. Wolf[i], Daiju Yamazaki[j], Alan Boobis[k]

[a] Health and Environmental Sciences Institute, Washington, DC, USA
[b] Albert Einstein College of Medicine, Bronx, NY, USA
[c] National Institute of Environmental Health Sciences, National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, USA
[d] Corteva Agriscience™, Rome, Italy
[e] US Food and Drug Administration, Center for Food Safety and Applied Nutrition, Silver Spring, MD, USA
[f] Department of Genetic Toxicology, Covance Laboratories Ltd, Harrogate, North Yorkshire, HG3 1PY, UK
[g] Universidade de Lisboa and iMED.IL, Portugal
[h] Corteva Agriscience™, Newark, DE, USA
[i] Syngenta, Raleigh-Durham, NC, USA
[j] National Institutes of Health Sciences, Kanagawa, Japan
[k] Imperial College London, London, UK

ARTICLE INFO

Keywords:
New approach methodologies
Evaluation framework
Risk assessment
Human health

ABSTRACT

The need to develop new tools and increase capacity to test pharmaceuticals and other chemicals for potential adverse impacts on human health and the environment is an active area of development. Much of this activity was sparked by two reports from the US National Research Council (NRC) of the National Academies of Sciences, Toxicity Testing in the Twenty-first Century: A Vision and a Strategy (2007) and Science and Decisions: Advancing Risk Assessment (2009), both of which advocated for "science-informed decision-making" in the field of human health risk assessment. The response to these challenges for a "paradigm shift" toward using new approach methodologies (NAMS) for safety assessment has resulted in an explosion of initiatives by numerous organizations, but, for the most part, these have been carried out independently and are not coordinated in any meaningful way. To help remedy this situation, a framework that presents a consistent set of criteria, universal across initiatives, to evaluate a NAM's fit-for-purpose was developed by a multi-stakeholder group of industry, academic, and regulatory experts. The goal of this framework is to support greater consistency across existing and future initiatives by providing a structure to collect relevant information to build confidence that will accelerate, facilitate and encourage development of new NAMs that can ultimately be used within the appropriate regulatory contexts. In addition, this framework provides a systematic approach to evaluate the currently-available NAMs and determine their suitability for potential regulatory application. This 3-step evaluation framework along with the demonstrated application with case studies, will help build confidence in the scientific understanding of these methods and their value for chemical assessment and regulatory decision-making.

1. Introduction

For well over a decade, efforts have been underway in academic, industry and government institutions to develop *in silico*, *in chemico* and *in vitro* methods to assess the mammalian and ecological toxicity potential of pharmaceuticals and other chemicals. More recently, such efforts have accelerated, sparked in the United States of America by the National Academy of Sciences (NAS) publications: Toxicity Testing in the Twenty-first Century: A Vision and a Strategy (2007) and Science and Decisions: Advancing Risk Assessment (2009), both of which advocated for improvement in "science-informed decision-making" for safety evaluation and risk assessment (National Research Council, 2007, 2009). In Europe, the Registration, Evaluation, Authorization and restriction of Chemicals (REACH) Regulation (EC No, 1907/2006)

* Corresponding author.
E-mail address: sparish@hesiglobal.org (S.T. Parish).

https://doi.org/10.1016/j.yrtph.2020.104592
Received 29 May 2019; Received in revised form 15 January 2020; Accepted 27 January 2020
Available online 01 February 2020

called for the use of NAMs where suitable, in assessing the toxicity of industrial chemicals (European Parliament and Council of the European Union, 2006). In the United States, the recently updated Toxic Substances Control Act specifies that in fulfillment of the law there must be effort made to reduce testing in vertebrate animals and implement NAMs (Lautenberg Chemical Safety Act), and the 2018 Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States represents the consensus of sixteen federal agencies on the development and application of new human-relevant testing strategies (EPA, 2016a; Interagency Coordinating Committee on the Validation of Alternative Methods, 2018). Therefore, creating and implementing NAMs that are effective and reliable for evaluating chemical safety is not only important for reducing vertebrate animal use but scientifically defensible *in vitro* assays are also legally mandated.

Building on the International Program on Chemical Safety (IPCS)/ International Life Sciences Institute (ILSI) mode of action/human relevancy (MOA/HR) framework (Boobis et al., 2006, 2008 Sonich-Mullin et al., 2001; Meek et al., 2003; Seed et al., 2005), Ankley et al. expanded the concept to incorporate the needs for ecologically relevant assessment endpoints in a publication describing Adverse Outcome Pathways (AOPs) (Ankley et al., 2010). This has provided a framework on which NAMs and technologies could be based and integrated into the previous MOA/HR framework for assessing both ecotoxicological and human health effects (Meek et al., 2014). The AOP concept now forms the foundation of many of the Organization for Economic Co-operation and Development (OECD) activities in the development of NAMs.

In the pharmaceutical sector, animal studies are not expected to identify every possible human adverse event but to provide guidance for deriving the maximum recommended starting dose (MRSD) for first-in-human (FIH) clinical trials as well as assessing the likelihood of organ toxicities and carcinogenicity. There have been very few, but notable, severe adverse outcomes in Phase I clinical trials, showing that animal studies contribute positively to identifying a safe starting dose (Suntharalingam et al., 2006; Bonini and Rasi, 2016). Nonetheless, there are examples of effects that were not predicted from the animal studies, highlighting that whole animal studies also have limitations on their predictivity in selected situations. Since there continues to be adverse events surfacing only in Phase III trials or post-market, even limited human *in vivo* data do not always allow accurate prediction of potential toxicity in the larger population. Hence, improvement is needed in predicting adverse events, and new approach methodologies (NAMs) have great potential. For many years, the pharmaceutical industry has used a discovery toxicology approach that incorporates addressing safety early in the pipeline to decrease attrition due to unanticipated toxicity identified only later in development. In 2000, Olson et al. assessed the predictive power of animal studies in relation to human toxicities observed in FIH or later clinical trials (Olson et al., 2000). In their report of the results of a multinational pharmaceutical company survey and the outcome of a multi-stakeholder workshop, the authors concluded that the true positive human toxicities concordance was 71% for rodent and nonrodent species, combined, and with nonrodents alone being predictive for 63% of human toxicities and rodents alone for 43% (Olson et al., 2000), thus leaving room for improvement.

In multiple regulated industry sectors at the global level, assessment frameworks for active ingredients and their associated final products, along with societal and regulatory initiatives to minimize or curtail animal use for toxicity testing, have been driving changes in safety characterization. Indeed, regulations such as those promulgated in the 7th Amendment to the European Union Cosmetics Directive (European Parliament and Council of the European Union, 2003) have placed an outright ban on animal testing for cosmetic products and their ingredients. Furthermore, some mechanisms that lead to adverse effects in animals have been shown to be irrelevant for humans, as for example, renal tumors in male rats induced by d-limonene as a consequence of binding to $\alpha 2_U$-globulin (Flamm and Lehman-McKeeman, 1991; Swenberg and Lehman-McKeeman, 1999). The response to these challenges is a call for a “paradigm shift” toward new, human relevant methods for risk assessment, which has resulted in a flurry of initiatives by numerous organizations. NAMs are envisioned to facilitate the replacement of animal testing with combinations of predictive *in silico* models (e.g., of human exposure structure-related toxicities, *in vitro* assays (e.g., human cell-based systems), and computational models of external and internal exposure (e.g., *in vitro* to *in vivo* extrapolation (IVIVE) and US EPA's ExpoCast) (European Chemicals Agency, 2016, 2017). These methods show promise for improving the speed and accuracy of chemical data needed for risk assessment. As with any method, whether traditional testing or new, it is necessary to determine its suitability for risk assessment and the associated decision contexts by ensuring that it meets the criteria associated with the intended use.

In light of the ongoing development of NAMs, initially occurring in response to animal welfare issues and represented by the 3Rs (reduction, refinement and replacement of animal methods), formal mechanisms for evaluation of proposed methods were established in the United States and Europe to provide scientific confidence and acceptance that a NAM was fit-for-purpose for its intended use, a key example of which is OECD Guidance Document (GD) 34 (OECD, 2005). A number of these new testing methods, for example assays for skin corrosion and serious eye damage or eye irritation (OECD, 2004; OECD, 2017a), have been successfully validated using the principles in GD34, and are accepted by regulatory authorities in predicting these toxicity endpoints. While the rate of development of new test methods has increased dramatically, the traditional time-consuming validation process does not enable decisions on the fitness of the new methods to match the pace at which they are being developed. Furthermore, NAMs, mainly *in vitro*, are being developed by many independent research groups, contract research organizations, and companies, in many countries, and with different funding streams. In addition, several large international research programs are attempting to facilitate cooperation among the various entities. In general, however, there is no overarching coordination of these different efforts in spite of their similar overall goals, resulting in unintended consequences, which include duplication of effort and limited input by the scientific community. These pitfalls and issues can result in lack of confidence in the methods and in significant delays in the implementation of new, possibly better, methods for assessing the potential for adverse effects. In addition, as the AOP concept has been implemented in method development, current evaluation procedures against apical effects in experimental animals are becoming increasingly less relevant (Interagency Coordinating Committee on the Validation of Alternative Methods, 2018; Piersma et al., 2018). Thus, a fit-for-purpose evaluation process must be able to match the nature and speed of scientific advancement while building confidence in the acceptability of NAMs. Hence, new approaches to establish confidence in NAMs are being developed, based on human biology and mechanistic relevance (i.e., AOP understanding), rather than an empirical approach based on apical effects in animal studies, for evaluating performance.

Therefore, the solution is to gain agreement on the objectives and approaches for evaluating and establishing credibility of NAMs, which would provide confidence to regulators, decision-makers, and the public, and lead to more rapid implementation. To that end, the Health and Environmental Sciences Institute (HESI) launched a project to establish a framework to guide in the evaluation of NAMs for human safety risk assessment based in the foundation of their context-of-use. The recommendations presented in this study do not constitute regulatory guidance and are not meant to supersede or supplant any existing regulatory policy, or address how NAMs could be implemented. Key aspects from OECD GD34 and other established validation principles (OECD, 2005; OECD, 2014), such as performance standards, acceptance criteria, peer review, statistical evaluation, etc. are used as a

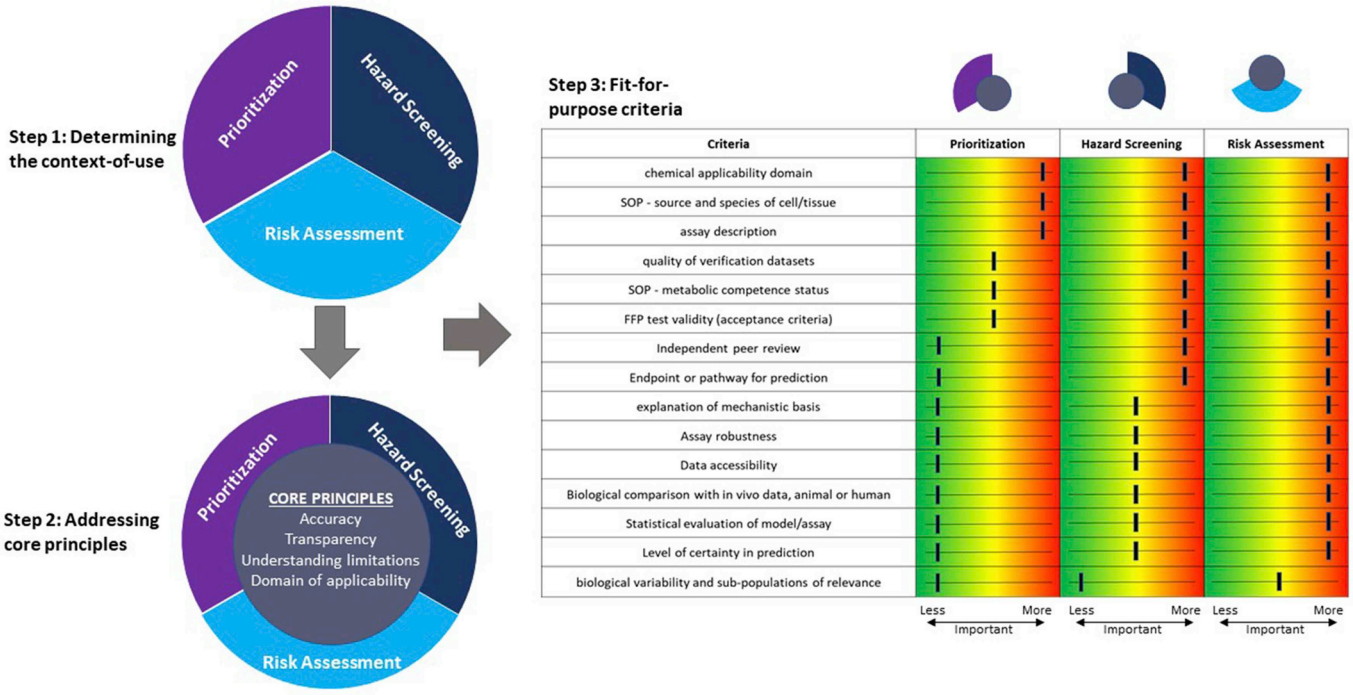


Fig. 1. The schematic represents the 3 steps of the framework. The first step (step 1) is to determine which of the 3 main contexts-of-use a NAM will be utilized (prioritization, hazard screening, or risk assessment). Step 2 is a set of core principles, irrespective of the context-of-use, that must to be addressed. Step 3 lays out criteria that will vary in their level of importance based on the context-of-use. While this schematic shows a default level of importance for each criterion, the importance of each should be considered on a sliding scale that will be dictated by various other factors (i.e. regulatory landscape, internal and external decisions) within a specific context-of-use.

foundation for the evaluation framework tool proposed here. This framework provides value in contextualizing the importance of those derived criteria, which were initially developed to provide guidance for regulatory or decision-making purposes, in a manner that orients NAM evaluations around higher vs. lower importance as driven by the context-of-use. The emphasis is on ensuring that a NAM is fit for its intended purpose, as determined by problem formulation. It is anticipated that the recommendations developed here may catalyze greater consistency of approach and implementation across a broad base of stakeholders. This public-private partnership, comprised of a global group of stakeholders from the industry, academic, and regulatory sectors, set out to develop a problem formulation strategy on the appropriate evaluation of NAMs based on the criteria needed to establish confidence in their performance.

2. The Framework

2.1. Rationale

As the development of NAMs continues at an increasing rate, there is a need for a framework that enables NAM assessment to keep up with the pace of development while ensuring they are fit-for-purpose, complementary to previous and existing efforts, and build upon those concepts and criteria where consensus has already been established. Within the present proposed framework, the information highlighted in OECD GD 34 and 211 (OECD, 2005; OECD, 2014) provide the criteria and accompanying definitions that facilitated the establishment of the core principles and criteria presented here. OECD GD 211 acknowledges that the information presented with the establishment of a NAM could vary in the level of detail, as based on the context-of-use and stage of development of a method. However, OECD GD 211 does not provide practical specifics on how these differences could be captured and portrayed as the context-of-uses change. The present framework extends well beyond the OECD guidance and does not duplicate or replace these previous efforts but rather complements and builds upon them. Its value resides in that it takes the information derived from those previous efforts and orients the evaluation of a NAM around the context-of-use through the organization of information based on a hierarchy of importance so that it is transparently articulated how a specific NAM can be used for a specific purpose. In this way, the development and application of NAMs becomes a dynamic process, ensuring always confidence in the data obtained for decision-making.

2.2. Establishing a problem formulation

In order to establish a foundation, the initial step is crafting the problem formulation for the question to be addressed or regulatory determination to be made (Interagency Coordinating Committee on the Validation of Alternative Methods, 2018). Problem formulation is a systematic and iterative method that sets out to define and characterize the issue at hand by understanding its key components, and then carrying out those relevant activities critical to addressing the issue (Sauve-Ciencewicki et al., 2019). In general, as defined more precisely below, problem formulation entails the narrowing of levels of abstraction and generality with the progressive sharpening of concepts and refinement of scope.

Within the different chemical sectors and regulatory agencies, problem formulation has been recognized as the first and most important step in risk analysis (National Research Council, 2009; Embry et al., 2014; Solomon et al., 2016). Identifying the problem that needs to be addressed (i.e., regulatory requirements, data needs, and context of use), and the best course of action to move forward in that assessment, is vital towards providing a foundation to problem resolution. As the process moves forward, the fitness of a method is dependent on the purpose for which it is intended. In some cases, the same method may be used for multiple purposes, and thus it may need multiple levels of evaluation relative to each use. For example, methods have been developed to be applied at one or more stages in the product development

process to screen for indicators of toxicity, identify hazards, compare relative potencies, and to predict similar adversity through read-across (Strickland et al., 2018, 2019; Choksi et al., 2018; Daniel et al., 2018). NAMs are also being considered to inform safety decisions in the regulatory process for new and existing compounds. A formal problem formulation process is useful to clearly articulate the purpose for which the method is to be used and to address acceptable levels of uncertainty, depending on the use, to build confidence in that method and ensure uptake and application across regulatory and industry stakeholders. Problem formulation for evaluation of a NAM includes three discreet steps: 1) determining the context of use; 2) addressing core principles and 3) defining fit-for-purpose criteria.

2.3. Step 1: Determining the context-of-use

First, one must determine where the method will be used (Fig. 1, step 1): Is the method to be used for i) prioritization, ii) hazard screening or iii) in a risk assessment? These are broad categories, and the definitions provided are to serve a general understanding of each use as finer divisions may be appropriate. The assignment into one of the three categories will determine which set of criteria needs to be addressed and possibly met (Fig. 1, step 3). Furthermore, determining the suitability of a NAM for a specific context-of-use should not only be questioned at the beginning, but as results are gathered, whether that NAM was suitable to provide the necessary information to address the goals and objectives set out, or whether additional information or alternative NAMs also need to be incorporated.

i) Prioritization: An evaluation that yields a rank order on a list of potential chemicals of interest. This ordering could be based on multiple factors, but the ultimate outcome is to have a list where those at the top will move forward into screening before those further down.
ii) Hazard Screening: This stage could use *in silico* or *in vitro* approaches to evaluate chemicals for their potential hazards, as determined by their intended use.
iii) Risk Assessment: This is usually a quantitative evaluation that, in addition to identifying the potential hazards during screening, incorporates exposure and dose into the final determination on a chemical's safety to a particular population.

It also needs to be noted that this process does not necessarily require assignment of a method to all three stages to be considered successful. Depending on business needs or the needs of a risk manager, NAMs may be necessary for only one of the stages described above. For example, in determining which contaminants most urgently need remediation, it may be sufficient to use methods for prioritization without moving on to hazard screening or risk assessment.

2.4. Step 2: Addressing core principles

Once the determination is made of how a method will be used, the core set of principles, in no particular order, of i) accuracy (e.g., true/false positive and negative detection), ii) transparency, iii) understanding limitations, and iv) domain of applicability, will need to be addressed (see Fig. 1: Step 2).

i) Accuracy: Often prediction models are required to consistently convert results from the NAM into *in vivo* toxicity predictions, and this may involve applying standardized data interpretation procedures to multiple tests in an integrated battery. Such prediction models permit objective comparison between the performance of a NAM and the *in vivo* study results. The prediction model should be presented, including data on the number of chemicals/substances used for the training of the model and those used for its development, and any overlap of these two sets should be made clear. It is therefore essential that method developers also develop the prediction model associated with each NAM prior to evaluation of the assay or integrated testing strategy.
ii) Transparency: The biological and technological basis of the NAM should be clearly described. This can lead to difficulties if there are implications for the protection of intellectual property. However, it is unlikely that widespread regulatory acceptance will be possible unless the basic principles of the method are known. Similarly, the algorithms and models used for data processing and extrapolation should be described, together with information on the chemicals/substances used for training the model and its performance in the assessment of those chemicals. Recently, OECD has provided guiding principles and best practices on accessing intellectual property in an effort to increase transparency, specifically when used in the development of a testing guideline program (OECD, 2019).
iii) Understanding limitations: An understanding of the key limitations of a NAM will facilitate appropriate use and interpretation in addressing the regulatory question. Knowing not only what a method can assess, but what the limitations are, will help ensure that the data generated are weighed appropriately. For example, metabolic competence is a major limitation for the majority of existing cell line-based NAMs (with the exception of certain whole cell cultures such as hepatocytes). Therefore, interpretation of negative results in these tests should be done with caution. Occasionally, additional tests may be required to confirm negatives from an assay conducted with a test system due to intrinsic limitations. The additional testing or battery testing may be a deterrent to move away from traditional tests, hindering the uptake of the newer methods. Clarity in testing strategies based on test limitations may instead favor acceptance and application, while limiting risks at the regulatory process level.
iv) Domain of applicability: This principle addresses the range a method will have, from which chemicals could be assessed (chemical applicability domain) to the biological species and pathway that could be evaluated. For example, test methods that are relevant to only water-soluble test materials (e.g., Cytosensor microphysiometer assay for determining ocular irritation) will not be applicable to highly lipophilic chemicals. It is just as important to characterize not only what falls within the range, but also what will fall outside the parameters of a method.

Addressing these first two steps of the process, i.e., determining the expected use(s) of a method coupled with addressing the core principles, lays the foundation for the problem formulation. One must also recognize that as more information is gained during the development of a method, this process is likely to require revision, through iterative reconsideration of the problem formulation. It will be necessary to revisit it often, ensuring that all those involved have a mutual understanding of the scopes and objectives (Sauve-Ciencewicki et al., 2019). Specifically, if a method is originally developed for one stage in the product life cycle (e.g., discovery vs. regulatory studies), and the question arises of a possible alternate utility, then the context of use changes and a new set of fit-for-purpose criteria might need to be addressed to establish confidence in its utility with those new requirements. These criteria under consideration, their definitions, and the level of importance constitute the third step in this framework.

2.5. Step 3 – Fit-for-purpose criteria

Once a context-of-use (Step 1) and the core principles are

determined and addressed (Step 2), the third step is to evaluate a list of criteria pertinent to assess the fit-for-purpose of a NAM (Fig. 1, step 3). The importance of each criterion considered within its context-of-use should be approached on a sliding-scale of importance. While Fig. 1, step 3 represents a default setting for each criterion within the three main context-of-uses, the level of importance should be adjusted to capture all the nuances within. These include, but are not limited to, any regulatory aspect and guidance and any internal decision-making processes. Ultimately, the goal is to orient the evaluation of a NAM to address the criteria in an order of importance for its specific context-of-use, as dictated under the problem formulation. Furthermore, these criteria could be applied either for a stand-alone test (e.g., a direct 1:1 replacement) or within a testing strategy, such as part of an IATA (integrated approach to testing and asessment) (OECD, 2005; OECD, 2016a).

2.6. Criteria definitions

Chemical applicability domain: It is important to understand performance of NAMs across diverse chemical classes. For example, a NAM may provide good predictions for only a small set of chemical classes relative to the universe of materials that need to be tested. If the NAM is limited in its application, it may only be relevant for specific situations for use in safety testing. In addition, assay systems that have limited metabolic capability may also have limited applicability and lead to false negative interpretations if metabolic activation is necessary. Furthermore, limitations might arise during assay development if only a limited number and classes of chemicals are used for the initial evaluation. Such a limitation in chemical space could pose a challenge for determining fit-for-purpose for applications or for providing the necessary certainty for a particular decision context. Therefore, every NAM must provide clarity on the domain of applicability of the assay, as well as clear interpretation criteria for positive and negative results. If a user wishes to apply a NAM to chemical classes outside the established applicability domain, the user should demonstrate that the assay applies to the new chemical class. Sometimes the proof will be categorical, e.g., a threshold for a result to be considered positive, but sometimes it will be quantitative, requiring weight of evidence considerations.

Standard operating procedure: An appropriate and clearly understandable and reproducible protocol is crucial for proper evaluation of a method and for gaining credibility and acceptance by regulatory agencies. The protocol should clearly indicate all study procedures, sources of cells including tissue and species, appropriate vehicle(s), positive and negative controls, assay acceptance criteria and data interpretation criteria. In addition, clear guidance documents and training materials that cover various critical technical aspects such as verification of cell type, number of cells, media, serum and additional components, incubation lengths and conditions, readout description, required equipment, and equipment calibration, should be described as completely as possible to facilitate replication of results (OECD, 2018a).

With regard to metabolic competence, most cell lines are severely deficient in, or even lack, enzymes for xenobiotic metabolism and other cellular processes (e.g., transport) and thus would not be considered fully metabolically competent. In many, but not necessarily all, cases, metabolic competence is necessary for relevant evaluation, thus information on the metabolic competence of the test system should be provided (i.e., OECD STA No. 280) (OECD, 2018b). If steps have been taken to address this, for example, by using an exogenous source of metabolizing enzymes or by genetically engineering cells to express enzymes for xenobiotic metabolism, then both qualitative and quantitative information on the performance of the metabolic system should be provided.

Method description: This criterion is met through written documentation of what an assay does and how it does it. This is separate from an SOP (previously described), which has the goal in providing a reproducible protocol on the execution of a method. The OECD Guidance Document 211 outlines the fields and definitions that would go into a method description, but the extent for that information will also be further dictated by the context of use for a NAM (OECD, 2014).

Quality of verification datasets: Positive and negative controls, in replicates, should always be included and well-characterized during method development, but a sufficient number of reference chemicals is key to establishing scientific confidence. In general, the selected reference chemical set should represent an adequate number of chemical classes to address the prediction targets (e.g., medical products, cosmetics, food additives and packaging, industrial chemicals, pesticides), and should be clearly related to the context of use for the assay. In addition, performance and limitations for application to specific chemical classes will need to be determined. Ideally, the reference data set should consist of chemicals clearly shown to be negative or positive in the assay or for the endpoint under consideration, with positive examples having sufficiently distributed potencies so that dose response can be adequately characterized.

Fit-for-purpose (FFP) test validity (acceptance criteria): Another key performance benchmark to judge the relevance of a method is to compare its prediction performance with *in vivo* results that are human-relevant. Although human *in vivo* data are extremely rare, critical evaluation of available rodent *in vivo* data to determine human relevance based on mode of action is possible and should be done before these comparisons are made (Boobis et al., 2006; Meek et al., 2003; Seed et al., 2005). The traditional approach has been to show reproducible results that are demonstrated across more than one laboratory, and where feasible, verification testing should be blind as to whether the chemical is positive or negative. Testing in multiple laboratories should ideally be performed but whenever this is not feasible or practical, performance-based criteria should be undertaken (see below). Such studies provide clarity on the performance of the selected NAM across diverse classification classes. If the sensitivity, specificity, and accuracy of the NAM are similar to or better than the *in vivo* values, it would support its relevance. Given that the ultimate objective is to predict human toxicity, ideally NAMs should be compared to human data, when available, that describe a key event(s) in an adverse outcome pathway. The relevant biological response, whether from human data or rodent assays, should be a component of a relevant pathway of toxicological concern that would result in a relevant response for a regulatory determination (Keller et al., 2012). It should not be necessary to recapitulate the apical endpoint but rather use an assay that represents one or more key events in a mode of action or adverse outcome pathway. Therefore, one could then see that this approach could enable more rapid, efficient, and relevant assessment of fit-for-purpose of NAMs. However, it requires a robust framework for judging the appropriateness of intermediate effects, which necessarily should include not only qualitative considerations but also potency and dose-response relationships. OECD is developing specific guidance on this that may be applicable, however the IPCS Human Relevance Framework can also inform the evaluation of these relationships (Boobis et al., 2006; Seed et al., 2005). Additionally, assay developers and regulators should work together to determine adequate criteria for acceptance for certain contexts of use.

Independent peer review: Depending on the purpose of the method,

it may require independent scientific peer review. This review could take on different forms, from publication in the scientific literature to a formal review under, for example, a Federal Advisory Committee, and will depend on the purpose of the method. Publications on the NAM's performance in high-quality peer-reviewed journals provide additional credibility for its validity. That credibility is also dictated by the extent either raw data or other details are described and shared within the publication process. In most cases, to gain acceptance for regulatory decision-making, the test method and supporting data would be reviewed by independent experts to determine the appropriateness of the NAM for the proposed purpose. NAMs that are considered reliable and relevant, for example, may be adopted to inform industry-based business decisions even before or during consideration for acceptance for use in official regulatory decisions.

Endpoint or pathway for prediction: This criterion addresses what is being measured, or what biological process is being assessed. This may be a molecular, cellular, tissue, or apical effect. A clear description of what is being assessed by the method should be provided, as well as, how well there is a plausible linkage between the effect measured and how close that key event is to the apical effect of concern.

Explanation of mechanistic basis: Recently, there has been heightened interest in developing an understanding of underlying biological pathways (AOPs), perturbation of which may lead to toxicological and pathological responses. The knowledge about these pathways has led to development of test methods that are designed to address specific key molecular and cellular events. Data from such NAMs would increase scientific confidence in the results and facilitate widespread acceptance of the method [e.g., skin sensitization tests focused on specific key events of the AOP] (OECD, 2016b). Current efforts are supporting the integration of toxicokinetics into *in vitro* evaluations of toxicodynamics. In addition, extrapolation from actual human exposure levels to *in vitro* concentrations and then back to the *in vivo* situation is critical for accurate interpretation for risk evaluation. Methods that enable *in vitro* to *in vivo* extrapolation (IVIVE) are necessary in order to accurately estimate relevant human exposures that correspond to observed *in vitro* bioactivity. IVIVE approaches should also be coupled with physiologically-based pharmacokinetic (PBPK) modelling to quantitatively bridge *in vitro* and *in vivo* data and to explore the key mechanisms dictating the pharmacokinetics. Combining *in vitro* methods with appropriate exposure data will improve applicability in a risk assessment framework. This would allow specific consideration with regard to route of exposure, target-specificity, and the potential for human extrapolation.

Assay robustness: The conventional idea describing this criterion is to assess both intra and inter-laboratory variability, which is typically addressed through multiple rounds of testing both within the same and across multiple laboratories. Typically, this requires a significant amount of time and resources. Increasingly, with rapidly changing science, there arise circumstances where inter-laboratory studies are not feasible due to a requirement for specialized equipment or expertise or the proprietary nature of the assay (e.g., evaluation of ocular irritants in PorCORA assay (Piehl et al., 2010), or skin sensitizers in the GARD assay (Johansson et al., 2013)). In such cases, or in the case of high-throughput screening (HTS) data generated by, for example, the Tox21 federal consortium, a performance-based evaluation could be considered (Judson et al., 2013). In this approach, the performance of a new test is validated against the results for previously validated tests for the same endpoint (e.g., estrogen receptor agonist and antagonist assays). Rather than employing multiple laboratories, the robustness of the assay is characterized using large sets of reference chemicals with well-defined activities in the existing methods. An example of an existing process is the FDA's qualification process, which decouples the analytical performance of an assay from the relevance, along with using laboratories that are well-trained in the type of assay being conducted.

Data Accessibility: In line with transparency, access to relevant, high quality (thoroughly curated), reference data is one of the most important factors for evaluation of NAMs. This access is instrumental in the ability to harmonize methods and to carry out any potential cross-comparative analysis. Initiatives such as the US EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) and the NICEATM Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/) are designed to facilitate access to curated datasets in a way that adheres to the FAIR principles, with data that should be findable, accessible (by both human and machine), interoperable, and reusable (Wilkinson et al., 2016).

Biological comparison with *in vivo* data (animal or human): In order for a method to provide information for use in a risk assessment, it will need to have relevance to the biology it is assessing. Comparison to *in vivo* data, either animal or human, is needed to gain acceptance, particularly in a regulatory setting. This does not mean one-to-one comparisons, where the results of a NAM are compared directly with the outcome of a cancer bioassay, or a reproductive toxicity study. Rather, information should be provided on the reliability of the method when used as envisaged to provide assurance on the potential of a chemical to produce such effects in the population. This could take a number of forms, such as comparison with known impacts of biological processes on human disease, e.g. inborn errors. Often, method assessment will be as part of an integrated testing and assessment strategy, rather than as a stand-alone method.

Statistical evaluation of model: Dealing with the collection, organization, analysis, and interpretation of the data set reflects how that sample is representative of a population. Determining the appropriate statistical approach and evaluation contributes to the scientific confidence that the results are predictive for a population.

Level of certainty in prediction: The data generated from a NAM has an associated level of confidence in its ability to be predictive for an endpoint or specific population. This criterion provides an understanding of what level of confidence a method operates within towards that predictive utility. In practice, this means that some methods will be more conservative (*i.e.,* yielding greater uncertainty and higher potential for false positives) by design for some applications than for others, for example for prioritization compared to risk assessment.

Biological variability and sub-populations of concern: As more data are generated that shed light on variability within populations, this criterion on how well a method captures differences within and between sub-populations will become increasingly applicable. Together with criteria for assessing the biological relevance of a method, this criterion on understanding how well the method can be utilized for assessing human variability will be instrumental in determining its applicability to specific sub-populations.

Coupling together the level of importance along with the understanding of each criterion, under a specific context-of-use, either a developer or an end-user can determine which criteria should be more important over others in the evaluation of a NAM. As the context of use changes (changes in step 1), or even within a context as additional factors needs to be incorporated (e.g. differences across regulatory considerations) so will the importance of each criterion (Fig. 1, Step 3). The three steps presented constitute the framework that either a developer or end-user could use as a tool to assess the fit-for-purpose of a method. The next section will illustrate the application of this framework through case studies.

3. Case studies

The four case studies presented below are meant to highlight the process by which one could follow the steps outlined in the framework, providing information on the various core principles and criteria. They were selected as they provide a broad range of NAMs at various stages in their respective development. Ultimately, an end-product would be a high-level summary of the specific NAM, illustrating side-by-side those crtieria of higher importance versus lower, and whether or not information is known for a specific criterion as it relates to that case example. For those criteria where information is not known, it is indicated as 'not provided" to demonstrate that there was no information captured to address that specific criterion within that example. Ultimately, a decision regarding the implementation and use of a NAM within a safety assessment will depend on various factors (e.g., weight of evidence) and will be determined by the evaluator or end-user. This process is intended to be a transparent and flexible tool that captures information on key criteria, while identifying any potential gaps that remain in a NAM's development and evaluation.

3.1. Case study #1: Ocular irritation tests

The evaluation of NAMs for ocular irritation testing is used to illustrate key concepts of applying the framework. Three tests, used in a weight-of-evidence approach, were considered: Bovine Corneal Opacity and Permeability test (BCOP), EpiOcular™ assay (EO), and Cytosensor Microphysiometer assay (CM). Assessment of acute eye irritation potential is required for chemicals prior to transportation and commercialization. Until recently, the *in vivo* Draize rabbit ocular irritation test was the only accepted test for the determination of the full range of irritation potential (severe irritants, irritants and non-irritants) by regulatory agencies worldwide. However, recently several *in vitro* test methods have gained regulatory acceptance for the identification of severe ocular irritants and ocular non-irritants (OECD, 2017b; McConnell et al., 1992; OECD, 2018c). All three methods are now considered appropriate for use at various stages of product development such as early in screening at active ingredient discovery stage, screening during product formulation development, and to inform determinations for classification and labelling. It should also be noted that to measure potency categories required for classification and labelling then one would need to address criteria further down the list (i.e. including those that are indicated in the yellow range for hazard screening), positive/negative hazard identification, then the bar would be much lower.

Step 1: Determining use: hazard screening and/or labeling of potential ocular irritants and non-irritants

Step 2: Defining core principles

i) Accuracy: The balanced accuracy for BCOP, EO and CM for identifying chemicals range from 69 to 85% compared to *in vivo* rabbit Draize eye irritation test results.

ii) Transparency: The prediction models and algorithms for determining performance of the negative and positive controls as well as test materials have been clearly defined for BCOP, CM and EO assays (OECD Test Guidelines 437 and 492) (OECD, 2017b; McConnell et al., 1992; OECD, 2018c). Isolated bovine corneas are obtained as a by-product from freshly slaughtered animals, while EO tissue and L929 cells (used for the CM assay) are commercially available. The procedures for these methods are transparent and published. Currently, only a very few contract research laboratories have the instrumentation used for CM, which poses a significant challenge to testing chemicals using the CM approach. Likewise, there is only a single producer of the opacimeter required for the BCOP assay; therefore, availability of specialized instrumentation should be considered in determining testing requirements.

iii) Understanding key limitations: The *in vivo* Draize eye test quantifies ocular irritation by measuring injury to ocular tissue and reversibility of effects within 21 days after initial exposure. The BCOP, EO and CM only measure injury to the tissue, not recovery. As a result, these assays provide poor predictions for differentiating between moderate and mild irritants. Furthermore, the cells and tissues used for each of these methods possess only a portion of the anatomical complexity present in human or rabbit eye, which, in general, over-predict chemical-mediated ocular irritation responses (reviewed in respective Test Guidelines).

iv) Domain of applicability: The chemical applicability has been established and provided in each Test Guideline. Additionally, these Test Guidelines are applicable to single substances and mixtures, and to solids, liquids, semi-solids and waxes. The liquids may be aqueous or non-aqueous, and solids may be soluble or insoluble in water. Whenever possible, solids should be ground to a fine powder before application, and no other pretreatment of the sample is required. Gases and aerosols have not been assessed in a validation study. While it is conceivable that these can be tested using Reconstructed human Cornea-like Epithelium (RhCE) technology, the current Test Guidelines do not apply to the testing of gases and aerosols.

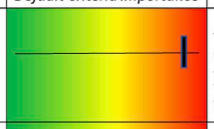
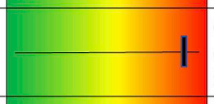
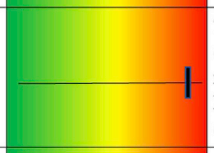
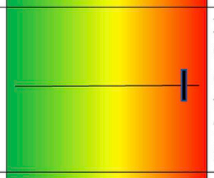
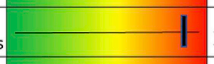
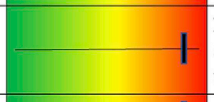
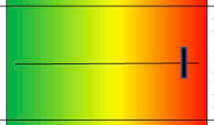
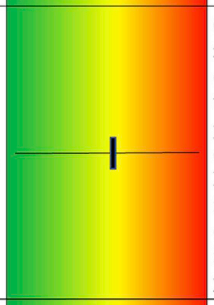
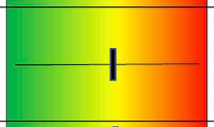
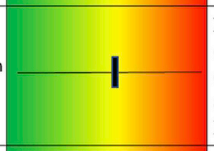
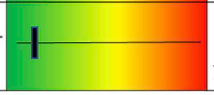
Step 3: Fitness-for-purpose criteria for hazard screening and/or labeling

The Table 1 below indicates the degree to which a specific criterion has been addressed within this case study, using the default setting of a hazard screen (Fig. 1) as to the importance of each criterion.

In this case study, the objective is a "Yes/No" answer as to whether a chemical is likely to be irritating to the eyes, so some of the criteria were considered less important than others. Hence, detailed mechanistic information on all of the steps between exposure and irritation was not considered essential, as the processes that are assessed, have been established as causal in the outcome. Assay robustness, as opposed to test validity (i.e. sensitivity and specificity) is not as important when the method is used in one laboratory for the intended purpose. Data accessibility is considered less important, as some assays use proprietary technology. However, information on assay performance should be accessible, and would be covered under other criteria. Biological comparison with *in vivo* data, animal or human is of less importance, as this information would already be available for the verification chemicals. In some circumstances, such as for testing cosmetics in Europe, a direct comparison would not be possible (at least with animal data). Formal statistical evaluation of model/assay is not of high importance, for the reasons explained above (no direct comparison with *in vivo* effects). The level of certainty in prediction is not as important, as long a minimum predictivity is achieved, as determined by verification against a reference set of chemicals.

Based on the fit-for-purpose criteria determined to be more important for a hazard screen, the methods address those standards. In this case, as additional criteria are addressed, although of lesser importance to the intended use of a hazard screen, may indicate its potential application in other contexts with further development. If the determination is made to use in a different context, then the importance of each criteria will change and need to be addressed as it pertains to that specific context-of-use.

Table 1
Hazard screening criteria details for ocular irritation.

Criteria	Hazard Screening *Default Criteria Importance*	Ocular Irritation
chemical applicability domain		A wide range of chemicals, covering a large variety of chemical types, chemical classes, molecular weights, LogPs, chemical structures, etc., have been tested in the evaluation study underlying these Test Guidelines
SOP - source and species of cell/tissue		Guidance for study conduct and interpretation is established for these methods and details can be found in the test guidelines
assay description		The *in vitro* and *ex vivo* assays provide a decision tree determination of eye irritation potential through the measurement of cytotoxicity under the United States Environmental Protection Agency (U.S. EPA) Office of Pesticide Programs (OPP).
quality of verification datasets		European Chemical Industry Ecology and Toxicology Centre (ECETOC) developed a reference databank containing 55 chemicals with *in vivo* eye irritation data that were generated using tests conforming to OECD Test Guideline 405. Respective OECD test guidelines include defined positive and negative controls
SOP - metabolic competence status		Limited metabolic capability for all three assays
FFP test validity (acceptance criteria)		Each of the three test methods have prescribed acceptability range for negative and positive controls to determine validity of the study
Independent peer review		All three assays were evaluated by ICCVAM, ECVAM
Endpoint or pathway for prediction		BCOP, EO and CM assays are based on quantifying chemical-mediated cytotoxicity upon direct exposure to isolated bovine corneal tissue, 3-D human keratinocytes and 2-D L929 cells, respectively.
explanation of mechanistic basis		Chemical-induced serious eye damage/eye irritation manifested *in vivo* mainly by corneal opacity, iritis, conjunctival redness and/or conjunctival chemosis, is the result of a cascade of events beginning with the penetration of the chemical through the cornea and/or conjunctiva and production of damage to the cells. However, it has been shown that cytotoxicity plays an important, if not the primary, mechanistic role in determining the overall serious eye damage/eye irritation response of a chemical regardless of the physicochemical processes underlying tissue damage
Assay robustness		The BCOP, CM and EO assays have undergone inter-laboratory testing. Results generated in these studies demonstrated good reproducibility and transferability of protocols within- and between laboratories
Data accessibility		All data used for validation are publicly available
Biological comparison with in vivo data, animal or human		BCOP, EO and CM assays quantify chemical-mediated cytotoxicity, which is considered to be one of the common endpoints leading to ocular irritation. These assays utilize isolated bovine corneal tissue, 3-D human keratinocytes and 2-D L929 cells
Statistical evaluation of model/assay		Not provided
Level of certainty in prediction		BCOP, EO and CM have accuracies of 82, 80, and 85%, respectively.
biological variability and sub-populations of relevance		Not provided
	Less ← → More Important	

3.2. Case study #2: Dermal sensitization

To expand on the evaluation of NAMs and continue to illustrate important concepts in the application of the framework, this case study describes an approach for a hazard screening of dermal sensitization using *in vitro* tests. Skin sensitization is one of the few toxicological endpoints for which an adverse outcome pathway (AOP) has been established and formally described (OECD, 2012a; OECD, 2012b). The AOP consists of four key events (KEs) starting with covalent binding of sensitizers to dermal proteins (KE1, haptenation; postulated to be the molecular initiating event (MIE)), followed by KE2, the activation of epidermal keratinocytes, KE3, the maturation and mobilization of Langerhans cells and dermal dendritic cells (DC), and KE4, the DC-mediated antigen presentation to naïve T-cells and activation and proliferation of allergen specific T-cells (OECD, 2016b). Mechanistic knowledge of the KEs has informed the development of a number of NAMs. As none of the NAMs alone covers the complexity of dermal sensitization, an IATA is necessary to adequately identify the potential hazard and related potency assessment for dermal sensitization (OECD, 2016b; OECD, 2012a; OECD, 2012b; Sauer et al., 2016). The IATA makes use of three methods as formally described by OECD, direct peptide reactivity assay (DPRA), ARE-Nrf2 Luciferase test method (KeratinoSens™ assay), and the Human Cell Line Activation test (h-CLAT) (OECD, 2012a; OECD, 2012b; Bauch et al., 2012; Urbisch et al., 2015). All three methods have been developed mainly for hazard identification purposes, however, information from these methods can be used for potency prediction when used together with other available information in an IATA format (Kleinstreuer et al., 2018a).

Step 1: Determining use: Hazard screening of dermal sensitization
Step 2: Defining core principles

i) Accuracy: The accuracy in predicting responses in the LLNA is 80% for a set of 157 chemicals (Kleinstreuer et al., 2018a)
ii) Transparency: The test methods for the DPRA, KeratinoSens™ and h-CLAT are described in OECD TG 442C, 442D and 442E respectively (OECD, 2015a; OECD, 2015b; OECD, 2018d). The three SOPs are available as DB-ALM protocol numbers 154, 155 and 158 (OECD, 2015a; OECD, 2018d; EURL ECVAM, 2019a; EURL ECVAM, 2019b; EURL ECVAM, 2019c).
iii) Understanding key limitations: The test methods described in this case study cannot be used on their own for potency predictions. As these methods possess limited metabolic capability, pro-haptens and pre-haptens with slower oxidation rates may provide false negative results. In the DPRA, substances that promote oxidation of peptides (without covalent interaction) may lead to false positive predictions. Substances that are highly cytotoxic or those that interfere with luciferase tend to result in false predictions in the KeratinoSens™ assay whereas pre or pro-haptens as well as strongly fluorescing substances may lead to false prediction in the h-CLAT assay.
iv) Domain of applicability: The test method is applicable to the testing of mono-constituent organic substances soluble in one of the solvents prescribed in the SOP. Only limited information is available on the applicability of the method for multi-constituent substances and mixtures of known composition.

Step 3: Fit-for-purpose criteria for hazard screening

Table 2 below indicates the degree to which a specific criterion has been addressed within this case study.

The reasons that some of the criteria are considered less important versus others is explained above for the case study on eye irritation. Again, as other nuances within a hazard screen might be considered (e.g. internal decisions, addressing specific regulatory needs), some criteria may increase in their importance. In this specific case, based on those that are deemed be on the higher end of importance, this case study addresses those criteria. Additionally, it should also be noted that a specific criterion addressed for one context-of-use might not be adequately addressed for another (i.e. meeting the standard for a hazard screen might not be enough in a risk assessment).

3.3. Case study #3: Ion-channel mediated cardiac toxicity

Cardiotoxicity includes direct drug or chemical effects on the heart, or indirect effects due to thrombotic events or alterations in hemodynamic flow (Albini et al., 2010). Cardiotoxicity is one of the leading causes for drug failure, either during drug development or after a compound has been approved for therapy (Li et al., 2016).

Inhibition of hERG (human ether-a-go-go channel; now named KCNH2) encodes the inward rectifying voltage gated potassium channel in the heart (IKr), which is involved in cardiac repolarization. Inhibition of the hERG current causes prolongation of the QT interval, which could lead to the Torsades de Pointes (TdP) arrhythmia (Priest et al., 2008). Numerous structurally- and functionally-unrelated drugs block the hERG potassium channel resulting in lengthened ventricular action potentials and prolonged QT interval, making it imperative to investigate any new chemical for this potential adverse effect before human exposure (Haverkamp et al., 2000). A list of drugs associated with prolonged QT interval is available from www.crediblemeds.org. In the late 1980's and early 1990's, several drugs were withdrawn from the market due to their association with prolonged QT interval and associated risk of TdP (Gintant et al., 2016). In response, emphasis on detecting drug-induced prolonged QT interval resulted in the ICH guideline S7B [The non-clinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation)] (ICH, 2005). This case study describes the development of a method to carry out a hazard screen for potential cardiotoxicities.

Step 1: Determining Use: Hazard screening of cardiotoxicity induced by inhibition of the hERG channel
Step 2: Defining core principles

i) Accuracy: This model's accuracy is assessed across 28 drugs with known clinical outcomes with a range of 75–80% accuracy (Li et al., 2019).
ii) Transparency: The hERG assay has been an ICH guideline since 2005, and since then has been widely utilized for screening drug candidates. The cell line typically used in the assay is a CHO-hERG line that stably expresses human ERG potassium channels. The procedures for conducting the hERG assay are transparent and freely available (Li et al., 2019).
iii) Understanding key limitations: Although the hERG assay is highly sensitive to compounds that might cause TdP, there are concerns that the screen is too conservative resulting in false positives and has low accuracy in predicting actual clinical risk (Colatsky et al., 2016). The assay will detect only those compounds that cause arrhythmia by inhibiting the hERG channel.
iv) Domain of applicability: Since 2005, practically every compound in drug discovery has been subject to the hERG assay. The assay can be performed using high-throughput automated patch clamp methods allowing large numbers of compounds to be screened (Houtmann et al., 2017). According to the ICH–S7B Guideline, consideration of compounds within a similar chemical class should be used as reference compounds and be included in the integrated risk assessment.

Step 3: Fitness-for-purpose criteria for hazard screening

Table 3 below indicates he degree to which a specific criterion has been addressed within this case study.

Based on the fit-for-purpose criteria outlined in this framework, this method addresses the more important criteria for a hazard screening. This assay predicts QT prolongation but not TdP which is the cardiac concern. There are many promising compounds that have been abandoned that may, in fact, not cause TdP. The Comprehensive *In Vitro*

Table 2
Hazard screening criteria details for dermal sensitization (Alves et al., 2016; Hoffmann et al., 2018; Wang et al., 2017).

Criteria	Hazard Screening *Default Criteria Importance*	Dermal Sensitization
chemical applicability domain		DPRA, KeratinoSens and h-CLAT methods are applicable to chemicals with a variety of organic functional groups, reaction mechanisms, potency and physicochemical properties
SOP - source and species of cell/tissue		Guidance for study conduct and interpretation are established for these methods
assay description		These assays address Key Events (KE) 1, 2 and 3 of the skin sensitization AOP
quality of verification datasets		A few curated databases integrated with *in vivo* (including human data when available) and *in vitro* data facilitate the development and evaluation of IATAs for dermal sensitization) [58-61]
SOP - metabolic competence status		KeratinoSens has limited metabolic activity as they are HaCaT cell derived. DPRA is cell-free and hCLAT has no inherent metabolic activity
FFP test validity (acceptance criteria)		Acceptance criteria for both the run and the chemical results are well defined.
Independent peer review		The assays' validation underwent an EURL ECVAM coordinated independent peer-review.
Endpoint or pathway for prediction		The mechanisms leading to dermal sensitization have been summarized in the form of an AOP, with the molecular initiating event being covalent binding between electrophilic moieties of chemicals and nucleophilic centers in dermal proteins.
explanation of mechanistic basis		DPRA, KeratinoSens assay, and the h-CLAT evaluate the MIE, KE2, and KE3 of the dermal sensitization AOP.
Assay robustness		Individual assay ring-trials / performance data formally reviewed by ECVAM with within and between laboratory repeatability (where performed) in excess of 80% each predictive accuracy found to be akin to the LLNA or better
Data accessibility		All data used for the assessment of the assay are publicly available.
Biological comparison with in vivo data, animal or human		It is still not clear what specific aspects of haptenation *in vivo* may be most relevant such as amino acid selectivity, reaction rate and stability of protein conjugates.
Statistical evaluation of model/assay		Evaluation procedures are published in respective OECD test guidelines. Calculations for defined approaches are published [51, 61]
Level of certainty in prediction		Accuracy in predicting responses in the LLNA is 80% for a set of 157 chemicals
biological variability and sub-populations of relevance		Not provided

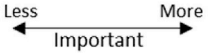
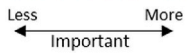


Table 3
Hazard screening criteria details for ion-channel mediated cardiac toxicity.

Criteria	Hazard Screening *Default Criteria Importance*	Ion-channel Mediated Cardiac Toxicity
chemical applicability domain		The assay can be performed using high-throughput automated patch clamp methods allowing large numbers of compounds to be screened [70]
SOP - source and species of cell/tissue		The ICH-S7B guideline outlines the general details of conducting the hERG assay
assay description		The hERG inhibition assay uses clonal cell lines to provide a sensitive measurement of a compounds' ability to inhibit hERG, which correlates with potential cardiotoxicity
quality of verification datasets		The quality and depth of hERG data sets are sufficient with many laboratories having a decade of historical control information
SOP - metabolic competence status		Very limited
FFP test validity (acceptance criteria)		ICH-S7B guideline includes general acceptance criteria prescribed and suggestions for quality assurance and the use of positive controls
Independent peer review		The ICH-S7B Guideline was developed by an ICH Expert Working Group, subject to consultation with the regulatory parties, and adopted in 2005 by the regulatory bodies of the European Union, Japan, and USA.
Endpoint or pathway for prediction		Blockade of the hERG potassium channel
explanation of mechanistic basis		Blockade of the hERG potassium channel is recognized as a predominant mechanism responsible for the drug-induced delayed repolarization linked to TdP
Assay robustness		The assay is currently performed globally and has good reproducibility and transferability within and between laboratories
Data accessibility		Not provided
Biological comparison with in vivo data, animal or human		Not provided
Statistical evaluation of model/assay		Not provided
Level of certainty in prediction		Not provided
biological variability and sub-populations of relevance		Not provided



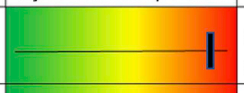
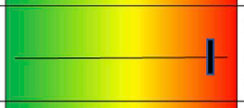
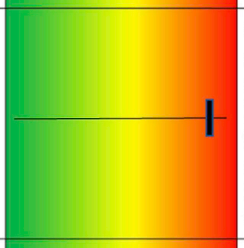
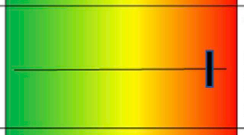
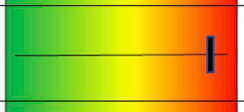
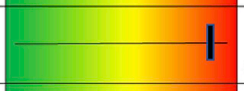
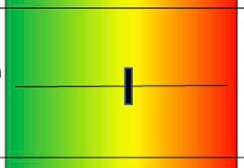
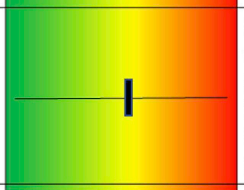
Proarrhythmia Assay (CiPA) is an initiative to combine more ion channel assays, *in silico* models, and *in vitro* myocytes to formulate a more complete picture of a compound's likelihood to cause TdP (CiPA, 2019).

3.4. Case study #4: Developmental vascularization

This case study describes the predictive signature using *in vitro* high-throughput screening (HTS) assays for identifying compounds that may interfere with developmental vascularization, i.e., putative vascular disruptor compounds (pVDCs). The model is based on the adverse outcome pathway (AOP) network for disruption of embryonic vascular development leading to adverse prenatal outcomes (Knudsen and Kleinstreuer, 2011), and herein this case study will apply the proposed criteria and describe how this predictive model could be used for screening environmental chemicals for developmental toxicity hazard via vascular disruption.

Step 1: Determining the Use: Prioritization and hazard screening of chemicals for impairment of developmental vascularization
Step 2: Defining core principles
i) Accuracy: Large sets of curated data for vascular disruption are

Table 4
Hazard screening criteria details for developmental vascularization.

Criteria	Hazard Screening *Default Criteria Importance*	Developmental Vascularization
chemical applicability domain		Constrained to chemicals already tested in the ToxCast HTS assays (or which could be screened in this way)
SOP - source and species of cell/tissue		Various human primary cells, cell-free biochemical assays, and cell lines. All part of ToxCast assay portfolio and described in publications.
assay description		Cardiovascular system formation is critical for all aspects of normal embryonic development, and environmental disruption results in diverse adverse prenatal outcomes. Multiple mammalian developmental toxicants specifically target vascular signaling molecules in HTS screens, and category formation based on bioactivity patterns can inform prioritization and screening of untested compounds for developmental toxicity.
quality of verification datasets		There are few known vascular disruptor compounds, but all such compounds identified via literature searches were identified as pVDCs, but biological variability remains, and more work is needed
SOP - metabolic competence status		very limited
FFP test validity (acceptance criteria)		The method is being used in certain limited contexts but extensive follow-up testing is underway to gain further acceptance
Independent peer review		The predictive signature for developmental vascularization, and the AOP network have been published in a series of peer-reviewed publications
Endpoint or pathway for prediction		AOP on embryonic vascular disruption described
explanation of mechanistic basis		The developmental vascular disruption AOP signature scores as part of the mechanistic data
Assay robustness		Tested in multiple labs across different functional vascular development methods (same set of 38 chemicals) to demonstrate robustness and predictive performance
Data accessibility		Data available online at https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data
Biological comparison with in vivo data, animal or human		Validation with known positive control thalidomide analogue, broader studies ongoing in transgenic zebrafish, human cell based tubulogenesis, and other complex cell culture models to test predictions across wide range of chemicals (Saili et al. 2019, manuscript in preparation).
Statistical evaluation of model/assay		Not provided
Level of certainty in prediction		Predictive signature is mapped to AOP for embryonic vascular disruption, however not all critical vascular signaling targets (e.g. cadherins) are represented due to lack of assay coverage to date. Uncertainty remains in the *in vitro* to *in vivo* extrapolation (toxicokinetic and exposure considerations), lack of metabolism in HTS systems.
biological variability and sub-populations of relevance		Not provided
	Less ◄ Important ► More	

not available, necessitating a reliance on individual reference chemicals and cross-validation using functional vascular development assays, e.g., in small model organisms such as zebrafish or human cell-based tubulogenesis assays. Reference VDCs, such as the thalidomide analogue 5HPP-33, were accurately predicted from both a hazard and a potency perspective (Kleinstreuer et al., 2013; Ellis-Hutchings et al., 2017). The pVDC predictions were further assessed across ten *in vitro* platforms from laboratories addressing different aspects of the vasculogenic/angiogenic cycle by testing 38 chemicals representing a range of VDC scores and various toxicity outcomes, including roughly equal numbers of predicted pVDCs and non-pVDCs that did or did not cause developmental toxicity. Overall, the pVDC signature prediction sensitivity and specificity were 89% and 80%, respectively; prediction accuracy was 87%, with greater predictivity of true positives (PPV 93%) compared to predictivity of true negatives (NPV 73%) (Saili et al., 2019; manuscript in preparation).

ii) Transparency: The principle of transparency and associated criteria of data and algorithm accessibility are highly applicable to the developmental vascularization AOP model, since this signature was developed based on publicly available data generated in the Tox21/ToxCast U.S. federal research program, and as such falls under a U.S. government mandate to be as open and transparent as possible. Those wishing to evaluate particular chemicals have access to all the source data (EPA, 2019) and the biological framework in which they are combined, resulting in an ability to prioritize chemicals for their potential vascular developmental hazard with respect to a large chemical library that includes known reference VDCs.

iii) Understanding key limitations: The pVDC AOP signature is currently based on a range of cellular and molecular targets from the Tox21/ToxCast screening program, largely in *in vitro* HTS assays that are commercially available. To apply this method to a particular chemical, the chemical must already be in the ToxCast library or be run in a suite of commercial assays. Further, this method could not be considered a stand-alone tool for developmental toxicity prioritization since it only addresses one key mechanism; however, it has high positive predictive value and could be used effectively in combination with other data as a screening tool. It could also be used in product development, to eliminate potential VDC liability.

iv) Domain of applicability: The predictive signature is applicable to chemicals amenable to HTS, and has been applied already to the ToxCast Phase I & Phase II chemical library.

Step 3: Fitness-for-purpose criteria for hazard screening

Table 4 below indicates the degree to which a specific criterion has been addressed within this case study.

Based on the fitness-for-purpose criteria determined more important for a hazard screen, this case study addresses those criteria in addition to some deemed less important. Work is underway to address the remaining criteria where information is currently not provided.

3.5. Application of these case studies to the context-of-use

Although these four case-studies present various applications within the context of a hazard screen, one can also see the potential opportunities to apply these NAMs in other contexts. Specifically, they all meet the criteria of high importance for a prioritization (i.e. Fig. 1, step 3 under prioritization), and identify gaps in criteria that are of high importance for application in a risk assessment, context (i.e. Fig. 1, step 3 under risk assessment). Furthermore, these case-studies are presented at various stages of their development, demonstrating that this framework could be utilized as a tool to continually ensure that those criteria of higher importance are met over others as dictated by the context-of-use of the NAM. Specifically, one would not need to wait until the end to evaluate a NAM but could use this framework to continually "check-in" during this iterative process and ensure focus is maintained to address criteria of higher importance. Finally, it must be stressed that even if all the pertinent criteria for a particular context are addressed, the decision regarding use or implementation of a NAM will be dependent on additional factors (i.e. fit-for-purpose, incorporated into a larger weight-of-evidence approach, exposure data, regulatory context), but that implementation aspect is beyond the scope of this framework. This framework is intended to identify and document foundational information that is necessary to increase confidence in the performance of a NAM under its intended use.

4. Discussion

As the need to develop NAMs will only continue to grow, building consensus around a framework to provide guidance on key considerations and components of a fit-for-purpose evaluation of those methods will only increase in importance. This framework, while oriented around human health, has principles and criteria outlined that could be envisioned to be applied to other taxa and their risk assessment. While the question on how to implement NAMs was beyond the scope as many additional components are needed to make a determination on when a NAM should be utilized, this framework is designed to systematically incorporate relevant information to aid in a NAM's evaluation as fit-for-purpose. This framework's value is that it takes criteria that have been synthesized and agreed upon by previous initiatives (OECD, 2005; OECD, 2014), and presents those criteria in a way that allows them to be prioritized on the level of importance as dictated within a specific context-of-use. This approach enables an evaluator to assure that the NAM is appropriate for its intended use and a researcher to construct a NAM with sufficient confidence that it can be used for its intended purpose. Furthermore, this framework does not replace or alter the conclusions derived by those previous efforts, but rather shows the broad applicability of those criteria for a variety of contexts-of-use. With the conscientious effort to design this framework by incorporating many of those established key aspects, this will go a long way towards international harmonization and ultimate acceptance. The adoption of consistent criteria and a transparent framework by NAM developers will mitigate the variability of different approaches and facilitate the regulatory acceptance of new NAMs at a global level. Through this 3-step process, one focuses the development and evaluation of a NAM towards those areas of greater importance, optimizing resource use and increasing confidence in the use and application of these methods within safety assessment. Scientific confidence in the applicability of a NAM is strengthened when consistent criteria are used to assess its reliability. A standard set of criteria for new method evaluation would speed the determination of whether a method is suitable for its intended application in making regulatory (or other) decisions.

In order to aid in the understanding of how this framework would be utilized, a series of case studies was presented. While the case studies were specific to hazard screening, they demonstrated that irrespective of the type of methodology (e.g. *in vitro*, *in silico*, computational), the same process could be employed to provide confidence in their evaluation for a hazard screen. In general, the criteria appeared appropriate to facilitate that evaluation. Furthermore, the same evaluation framework could be applied within the context of prioritization or risk assessment to visualize where criteria may have been met or are still missing. This framework is a tool to orient the evaluation towards criteria of higher importance, giving an additional piece of information to aid the end-user in making the final determination on a NAM's utilization. Case studies of NAMs that have been successfully implemented for their specific context-of-use have demonstrated that not all the criteria need to be addressed, but did sufficiently do so for criteria of higher importance. When the case study did not address a specific criterion, as indicated by "not provided", a determination would need

to be made as to whether that is deemed important for that specific context-of-use. Moving forward, further case studies for other contexts of use would help demonstrate the utility of the framework and promote its incorporation.

External to the case studies presented, opportunities for immediate application of this framework have been reflected across government agencies. US EPA's Office of Pesticide Programs'(EPA/OPP) efforts to modernize the acute toxicity "six-pack" studies and reduce animal testing is meant to reduce barriers for the incorporation of NAMs to animal testing and facilitate the use of OECD *in vitro* assays, including consideration of the globally harmonized system of classification and labeling (GHS) (Environmental Protection Agency, 2016). In collaboration with the National Toxicology Program's Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), industry, and non-governmental organizations, the US EPA is making significant progress towards the adoption of IATA and the reduction of the use of animals in acute toxicity testing and repeat dose studies where possible. The US EPA and NICEATM are working to develop a database of curated acute toxicity data from agrochemical products, including acute inhalation toxicity data that will be part of NTP's Integrated Chemical Environment (ICE) (National Toxicology Program, 2019). The resulting database will be used to assess the variability within and across studies, to develop read-across approaches, and to compare it with results from alternative approaches, such as the GHS additivity formula or *in vitro* studies. A large database of rodent acute oral toxicity data was curated by US EPA and NICEATM and used as the basis for a global collaboration to build predictive *in silico* models for acute oral systemic toxicity, targeted towards regulatory needs identified across federal agencies (Kleinstreuer et al., 2018b). Furthermore, as part of the US EPA's commitment to reduce animal use, a series of guidance documents have been released over the years. Specifically in 2013, the US EPA's Office of Pesticide Program released the guidance document "Guiding Principles for Data Requirements" (EPA, 2013) followed by "Process for Establishing and Implementing Alternative Approaches to Traditional *In Vivo* Acute Toxicity Studies" in 2016 (EPA, 2016b). Subsequent to an US EPA OPP 2012 guidance issuance document (EPA, 2012), the OECD released a "Guidance Document for Waiving or Bridging Acute Toxicity Tests" (OECD, 2017c). US EPA is currently conducting a pilot utilizing the GHS dose additive mixtures equation to reduce animal testing for pesticide product formulations. The goal of the pilot is to evaluate the utility and acceptability of the GHS dose additive mixtures equation as an alternative to animal oral and inhalation toxicity studies for pesticide product formulations. A recent document (EPA, 2018), highlighted US EPA efforts towards the use of OECD's test guidelines for *in vitro* studies for eye irritation, skin irritation and skin sensitization. Finally, the US FDA recently published its Predictive Toxicology Roadmap that described a path forward for the agency to incorporate NAMs in its regulatory framework (Food and Drug Administration, 2019). The US FDA is currently working on its implementation plan for working with its stakeholders to accomplish these goals. The US FDA, DARPA, and NCATS Partnership for developing *In Vitro* Microphysiological Systems was an example of how government regulatory and other scientists working with academic and industry partnerships can move the development of an important innovative technology forward more rapidly than any one group alone. All of these efforts combined will enable a more rapid acceptance of NAMs to inform health protective decisions for the approval and use of new products.

Although this framework contributes towards a harmonization of processes and fit-for-purpose criteria for the evaluation of NAMs, there are still areas that need to be addressed. Challenges continue to remain in the evaluation and interpretation of results across different regulatory agencies, who are also constrained by differing legislative mandates. There are, however, ongoing efforts to understand similarities and difference across regulatory geographies. The International Cooperation of Alternative Test Methods (ICATM) recently looked at testing requirements, specifically to skin sensitization testing, across seven countries or regions to inform their strategies for the acceptance and implementation of NAMs (Daniel et al., 2018). Convening workshops of international stakeholders, as was carried out in 2016 to address alternative approaches to inhalation toxicity testing (Clippinger et al., 2018), will only aid in the effort to harmonize the evaluation and interpretation of NAMs across global regulatory agencies. The criteria outlined in this framework provide a sound foundation and additional piece to these ongoing efforts, but they are still somewhat preliminary. There is a need for greater scientific consensus on the appropriate criteria and the level of detail required in the evaluation and interpretation of NAMs for specific purposes. Until this is achieved, it will remain difficult to implement them for direct application in regulatory or other systematic decision-making contexts. Additionally, in order to gain acceptance, there will be a continued need to demonstrate biological relevance. The reality, however, is that in some areas of toxicity, human data are limited, do not exist, or are unethical to obtain. This current gap is a further reason to move to a more biologically/AOP-based approach to determining relevance. This means that the appropriate comparison would not be with the apical endpoints observed in animal studies, but with the relevant biology responsible for adverse outcomes in humans. This is not a binary question, but one of dose-response relationships, when homeostasis is overwhelmed, and adaptation overcome.

The application of the framework presented herein should facilitate and encourage development and accelerate the appropriate application of NAMs and build confidence in the scientific understanding of these assays and their value for chemical and pharmaceutical assessment and regulatory decision-making. The methods outlined in this framework are demonstrated through case studies representing NAMs differing in their methodology (i.e. *in vitro*, *in silico*, computational approaches) and at different stages in their development, which provide several perspectives in evaluating the proposed framework in supporting acceptability of a NAM. General application of this framework could support harmonization and provide a foundation to facilitate and encourage development and accelerate application of NAMs while building confidence in their value for chemical and pharmaceutical assessment and regulatory decision-making.

Funding

This HESI scientific initiative is primarily supported by in-kind contributions (from public and private sector participants) of time, expertise, and experimental effort. These contributions are supplemented by direct funding (that largely supports program infrastructure and management) that was provided by HESI's corporate sponsors. A list of supporting organizations (public and private) is available at http://hesiglobal.org.

Declaration of competing interest

The authors 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

The authors gratefully acknowledge the input from the government, academic, and industry scientists of the HESI Framework for Intelligent Non-Animal Methods for Safety Assessment subcommittee for their contributions to this work.

References

Albini, A., Pennesi, G., Donatelli, F., Cammarota, R., De Flora, S., Noonan, D.M., 2010.

Cardiotoxicity of anticancer drugs: the need for cardio-oncology and cardio-oncological prevention. J. Natl. Cancer Inst. 102 (1), 14–25.
Alves, V.M., Capuzzi, S.J., Muratov, E., Braga, R.C., Thornton, T., Fourches, D., et al., 2016. QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. Green Chem. 18 (24), 6501–6515.
Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., et al., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29 (3), 730–741.
Bauch, C., Kolle, S.N., Ramirez, T., Eltze, T., Fabian, E., Mehling, A., et al., 2012. Putting the parts together: combining *in vitro* methods to test for skin sensitizing potentials. Regul. Toxicol. Pharmacol. 63 (3), 489–504.
Bonini, S., Rasi, G., 2016. First-in-human clinical trials - what we can learn from tragic failures? N. Engl. J. Med. 375 (18), 1788–1789.
Boobis, A.R., Cohen, S.M., Dellarco, V., McGregor, D., Meek, M.E., Vickers, C., et al., 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Crit. Rev. Toxicol. 36 (10), 781–792.
Boobis, A.R., Doe, J.E., Heinrich-Hirsch, B., Meek, M.E., Munn, S., Ruchirawat, M., et al., 2008. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. Crit. Rev. Toxicol. 38 (2), 87–96.
Choksi, N.Y., Truax, J., Layton, A., Matheson, J., Mattie, D., Varney, T., et al., 2018. United States regulatory requirements for skin and eye irritation testing. Cutan. Ocul. Toxicol. 1–58.
CiPA, 2019. CiPA work streams. [cited 12 February 2019]. http://cipaproject.org/about-cipa/#WorkStreams.
Clippinger, A.J., Allen, D., Jarabek, A.M., Corvaro, M., Gaca, M., Gehen, S., et al., 2018. Alternative approaches for acute inhalation toxicity testing to address global regulatory and non-regulatory data requirements: an international workshop report. Toxicol. Vitro 48, 53–70.
Colatsky, T., Fermini, B., Gintant, G., Pierson, J.B., Sager, P., Sekino, Y., et al., 2016. The comprehensive *in vitro* Proarrhythmia assay (CiPA) initiative - update on progress. J. Pharmacol. Toxicol. Methods 81, 15–20.
Daniel, A.B., Strickland, J., Allen, D., Casati, S., Zuang, V., Barroso, J., et al., 2018. International regulatory requirements for skin sensitization testing. Regul. Toxicol. Pharmacol. 95, 52–65.
Ellis-Hutchings, R.G., Settivari, R.S., McCoy, A.T., Kleinstreuer, N., Franzosa, J., Knudsen, T.B., et al., 2017. Embryonic vascular disruption adverse outcomes: linking high-throughput signaling signatures with functional consequences. Reprod. Toxicol. 71, 16–31.
Embry, M.R., Bachman, A.N., Bell, D.R., Boobis, A.R., Cohen, S.M., Dellarco, M., et al., 2014. Risk assessment in the 21st century: roadmap and matrix. Crit. Rev. Toxicol. 44 (Suppl. 3), 6–16.
Environmental Protection Agency, 2016. Letter to stakeholders on EPA Office of pesticide programs's goal to reduce animal testing from Jack E. Housenger, director Office of pesticide programs. [cited 26 February 2019]. https://www.regulations.gov/document?D=EPA-HQ-OPP-2016-0093-0003.
Epa, 2012. Guidance for waiving or bridging of mammalian acute toxicity tests for pesticides and pesticide products (acute oral, acute dermal, acute inhalation, primary eye, primary dermal, and dermal sensitization). [cited 12 February 2019]. https://www.epa.gov/sites/production/files/documents/acute-data-waiver-guidance.pdf.
Epa, 2013. Guiding principles for data requirements. [cited 12 February 2019]. https://www.epa.gov/sites/production/files/2016-01/documents/data-require-guide-principle.pdf.
EPA, 2016a. The Frank R. Lautenberg chemical safety for the 21st century Act. [cited 12 February 2019]. https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act.
EPA, 2016b. Process for establishing and implementing alternative approaches to traditional *in vivo* acute toxicity studies. [cited 12 February 2019]. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/process-establishing-implementing-alternative.
EPA, 2018. Interim science policy: use of alternative approaches for skin sensitization as a replacement for laboratory animal testing draft for public comment. [cited 12 February 2019]. https://www.regulations.gov/document?D=EPA-HQ-OPP-2016-0093-0090.
EPA, 2019. Exploring ToxCast data: downloadable data. [cited 12 February 2019]. https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data.
EURL ECVAM, 2019a. Direct peptide reactivity assay (DPRA) for skin sensitisation testing, DB-ALM protocol no. 154. [cited 12 February 2019]. https://ecvam-dbalm.jrc.ec.europa.eu/methods-and-protocols/topic/sensitisation-and-allergy/key/t_41.
EURL ECVAM, 2019b. KeratinoSens™, DB-ALM protocol no. 155. [cited 14 August 2019]. https://ecvam-dbalm.jrc.ec.europa.eu/methods-and-protocols/topic/sensitisation-and-allergy/key/t_41.
EURL ECVAM, 2019c. Human cell line activation test (h-CLAT), DB-ALM protocol no. 158. [cited 14 August 2019]. https://ecvam-dbalm.jrc.ec.europa.eu/methods-and-protocols/topic/sensitisation-and-allergy/key/t_41.
European Chemicals Agency, 2016. In: New Approach Methodologies in Regulatory Science: Proceedings of a Scientific Workshop (19–20 April 2016), [cited 26 February 2019]. https://echa.europa.eu/documents/10162/22816069/scientific_ws_proceedings_en.pdf.
European Chemicals Agency, 2017. Non-animal approaches: current status of regulatory applicability under the REACH, CLP and biocidal products regulations. [cited 26 February 2019]. https://echa.europa.eu/documents/10162/22931011/non_animal_approcches_en.pdf/87ebb68f-2038-f597-fc33-f4003e9e7d7d.
European Parliament and Council of the European Union, 2003. Directive 2003/15/EC of the European parliament and of the Council of 27 february 2003 amending Council directive 76/768/EEC on the approximation of the laws of the member States relating to cosmetic products. [cited 26 February 2019]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32003L0015.
European Parliament and Council of the European Union, 2006. Regulation (EC) No 1907/2006 of the European parliament and of the Council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European chemicals agency, amending directive 1999/45/EC and repealing Council regulation (EEC) No 793/93 and commission regulation (EC) No 1488/94 as well as Council directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. [cited 12 February 2019]. http://data.europa.eu/eli/reg/2006/1907/oj.
Flamm, W.G., Lehman-McKeeman, L.D., 1991. The human relevance of the renal tumor-inducing potential of d-limonene in male rats: implications for risk assessment. Regul. Toxicol. Pharmacol. 13 (1), 70–86.
Food and Drug Administration, 2019. FDA's predictive toxicology roadmap. [cited 26 February 2019]. https://www.fda.gov/scienceresearch/aboutscienceresearchatfda/ucm601090.htm.
Gintant, G., Sager, P.T., Stockbridge, N., 2016. Evolution of strategies to improve preclinical cardiac safety testing. Nat. Rev. Drug Discov. 15, 457.
Haverkamp, W., Breithardt, G., Camm, A.J., Janse, M.J., Rosen, M.R., Antzelevitch, C., et al., 2000. The potential for QT prolongation and pro-arrhythmia by non-anti-arrhythmic drugs: clinical and regulatory implications. Report on a Policy Conference of the European Society of Cardiology. Cardiovasc. Res. 47 (2), 219–233.
Hoffmann, S., Kleinstreuer, N., Alepee, N., Allen, D., Api, A.M., Ashikaga, T., et al., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. Crit. Rev. Toxicol. 48 (5), 344–358.
Houtmann, S., Schombert, B., Sanson, C., Partiseti, M., Bohme, G.A., 2017. Automated patch-clamp methods for the hERG cardiac potassium channel. Methods Mol. Biol. 1641, 187–199.
ICH, 2005. Guideline S7B: non-clinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation) by human pharmaceuticals. [cited 12 February 2019]. https://www.ich.org/products/guidelines/safety/safety-single/article/the-non-clinical-evaluation-of-the-potential-for-delayed-ventricular-repolarization-qt-interval-pro.html.
Interagency Coordinating Committee on the Validation of Alternative Methods, 2018. A strategic roadmap for establishing new approaches to evaluate the safety of chemicals and medical products in the United States. [cited 12 February 2019]. https://ntp.niehs.nih.gov/iccvam/docs/roadmap/iccvam_strategicroadmap_january2018_document_508.pdf.
Johansson, H., Albrekt, A.S., Borrebaeck, C.A., Lindstedt, M., 2013. The GARD assay for assessment of chemical skin sensitizers. Toxicol. Vitro 27 (3), 1163–1169.
Judson, R., Kavlock, R., Martin, M., Reif, D., Houck, K., Knudsen, T., et al., 2013. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. ALTEX 30 (1), 51–56.
Keller, D.A., Juberg, D.R., Catlin, N., Farland, W.H., Hess, F.G., Wolf, D.C., et al., 2012. Identification and characterization of adverse effects in 21st century toxicology. Toxicol. Sci. 126 (2), 291–297.
Kleinstreuer, N., Dix, D., Rountree, M., Baker, N., Sipes, N., Reif, D., et al., 2013. A computational model predicting disruption of blood vessel development. PLoS Comput. Biol. 9 (4) e1002996.
Kleinstreuer, N.C., Hoffmann, S., Alepee, N., Allen, D., Ashikaga, T., Casey, W., et al., 2018a. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches (*). Crit. Rev. Toxicol. 48 (5), 359–374.
Kleinstreuer, N.C., Karmaus, A.L., Mansouri, K., Allen, D.G., Fitzpatrick, J.M., Patlewicz, G., 2018b. Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation. Comput Toxicol 8, 21–24.
Knudsen, T.B., Kleinstreuer, N.C., 2011. Disruption of embryonic vascular development in predictive toxicology. Birth Defects Res C Embryo Today 93 (4), 312–323.
Li, Z., Dutta, S., Sheng, J., Tran, P.N., Wu, W., Colatsky, T., 2016. A temperature-dependent in silico model of the human ether-a-go-go-related (hERG) gene channel. J. Pharmacol. Toxicol. Methods 81, 233–239.
Li, Z., Ridder, B.J., Han, X., Wu, W.W., Sheng, J., Tran, P.N., et al., 2019. Assessment of an in silico mechanistic model for Proarrhythmia risk prediction under the CiPA initiative. Clin. Pharmacol. Ther. 105 (2), 466–475.
McConnell, H.M., Owicki, J.C., Parce, J.W., Miller, D.L., Baxter, G.T., Wada, H.G., et al., 1992. The cytosensor microphysiometer: biological applications of silicon technology. Science 257 (5078), 1906–1912.
Meek, M.E., Bucher, J.R., Cohen, S.M., Dellarco, V., Hill, R.N., Lehman-McKeeman, L.D., et al., 2003. A framework for human relevance analysis of information on carcinogenic modes of action. Crit. Rev. Toxicol. 33 (6), 591–653.
Meek, M.E., Boobis, A., Cote, I., Dellarco, V., Fotakis, G., Munn, S., et al., 2014. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. J. Appl. Toxicol. 34 (1), 1–18.
National Research Council, 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. The National Academies Press, Washington, DC.
National Research Council, 2009. Science and Decisions: Advancing Risk Assessment. The National Academies Press, Washington, DC.
National Toxicology Program, 2019. Integrated chemical environment (ICE). [cited 12 February 2019]. https://ice.ntp.niehs.nih.gov/.
NICEATM, 2018. NICEATM murine local lymph node assay (LLNA) database. [cited 12 February 2019]. https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/immunotoxicity/nonanimal/index.html#NICEATM-Murine-Local-Lymph-Node-Assay-LLNA-Database.
OECD, 2004. Test No. 431: *in vitro* skin corrosion: human skin model test. [cited 12 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264071148-en.
OECD, 2005. Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment: OECD series on testing and

assessment, number 34 (ENV/JM/MONO(2005)14). [cited 12 February 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono 14.
OECD, 2012a. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins Part 1: scientific evidence. Series on testing and assessment No. 168 (ENV/JM/MONO(2012)10/PART1). [cited 12 February 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2012)10/part1&doclanguage=en.
OECD, 2012b. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. Part 2: use of the AOP to develop chemical categories and integrated assessment and testing approaches series on testing and assessment No. 168 (ENV/JM/MONO(2012)10/PART2). [cited 12 February 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282012%2910/part2&doclanguage=en.
OECD, 2014. Guidance document for describing non-guideline *in vitro* test methods, OECD series on testing and assessment, No. 211 (ENV/JM/MONO(2014)35). [cited 14 January 2020]. https://www.oecd-ilibrary.org/environment/guidance-document-for-describing-non-guideline-in-vitro-test-methods_9789264274730-en.
OECD, 2015a. test No. 442C: *in chemico* skin sensitisation. [cited 12 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264229709-en.
OECD, 2015b. Test No. 442D: *in vitro* skin sensitisation. [cited 14 August 2019]. https://www.oecd-ilibrary.org/environment/test-no-442d-in-vitro-skin-sensitisation_9789264229822-en.
OECD, 2016a. Guidance document for the use of adverse outcome pathways in developing integrated approaches to testing and assessment (IATA): series on testing and assessment No. 260 (ENV/JM/MONO(2016)67). [cited 14 August 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2016)67/&doclanguage=en.
OECD, 2016b. Guidance document on the reporting of defined approaches and individual information sources to Be used within integrated approaches to testing and assessment (IATA) for skin sensitisation: series on testing & assessment No. 256 (ENV/JM/MONO(2016)29). [cited 12 February 2019]. https://one.oecd.org/document/ENV/JM/MONO(2016)29/en/pdf.
OECD, 2017a. Test No. 405: acute eye irritation/corrosion. [cited 12 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264185333-en.
OECD, 2017b. Test No. 437: bovine corneal opacity and permeability test method for identifying i) chemicals inducing serious eye damage and ii) chemicals not requiring classification for eye irritation or serious eye damage. [cited 12 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264203846-en.
OECD, 2017c. Guidance document on considerations for waiving or bridging of mammalian acute toxicity tests. [cited 12 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264274754-en.
OECD, 2018a. Guidance document on good *in vitro* method practices (GIVIMP): series on testing and assessment No. 286 (ENV/JM/MONO(2018)19). [cited 12 February 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2018)19&doclanguage=en.
OECD, 2018b. Guidance document on the determination of *in vitro* intrinsic clearance using cryopreserved hepatocytes (RT- HEP) or liver S9 sub-cellular fractions (RT-S9) from rainbow trout and extrapolation to *in vivo* intrinsic clearance series on testing and assessment. No. 280 (ENV/JM/MONO(2018)12). [cited 12 February 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2018)12&doclanguage=en.
OECD, 2018c. Test No. 492: Reconstructed human Cornea-like Epithelium (RhCE) test method for identifying chemicals not requiring classification and labelling for eye irritation or serious eye damage. [cited 26 February 2019]. https://www.oecd-ilibrary.org/content/publication/9789264242548-en.
OECD, 2018d. Test No. 442E: in vitro skin sensitisation. [cited 16 May 2019]. https://www.oecd.org/env/test-no-442e-in-vitro-skin-sensitisation-9789264264359-en.htm.
OECD, 2019. Guiding principles on good practices for the availability/distribution of protected elements in OECD test guidelines: series on testing and assessment No. 298 (ENV/JM/MONO(2019)14). [cited 2 August 2019]. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2019)14&docLanguage=En.
Olson, H., Betton, G., Robinson, D., Thomas, K., Monro, A., Kolaja, G., et al., 2000. Concordance of the toxicity of pharmaceuticals in humans and in animals. Regul. Toxicol. Pharmacol. 32 (1), 56–67.
Piehl, M., Gilotti, A., Donovan, A., DeGeorge, G., Cerven, D., 2010. Novel cultured porcine corneal irritancy assay with reversibility endpoint. Toxicol. Vitro 24 (1), 231–239.
Piersma, A., Burgdorf, T., Louekari, K., Desprez, B., Taalman, R., Landsiedel, R., et al., 2018. Workshop on acceleration of the validation and regulatory acceptance of alternative methods and implementation of testing strategies. Toxicol. Vitro 50, 62–74.
Priest, B.T., Bell, I.M., Garcia, M.L., 2008. Role of hERG potassium channel assays in drug development. Channels 2 (2), 87–93.
Sauer, U.G., Hill, E.H., Curren, R.D., Raabe, H.A., Kolle, S.N., Teubner, W., et al., 2016. Local tolerance testing under REACH: accepted non-animal methods are not on equal footing with animal tests. Altern Lab Anim 44 (3), 281–299.
Sauve-Ciencewicki, A., Davis, K.P., McDonald, J., Ramanarayanan, T., Raybould, A., Wolf, D.C., et al., 2019. A simple problem formulation framework to create the right solution to the right problem. Regul. Toxicol. Pharmacol. 101, 187–193.
Seed, J., Carney, E.W., Corley, R.A., Crofton, K.M., DeSesso, J.M., Foster, P.M., et al., 2005. Overview: using mode of action and life stage information to evaluate the human relevance of animal toxicity data. Crit. Rev. Toxicol. 35 (8–9), 664–672.
Solomon, K.R., Wilks, M.F., Bachman, A., Boobis, A., Moretto, A., Pastoor, T.P., et al., 2016. Problem formulation for risk assessment of combined exposures to chemicals and other stressors in humans. Crit. Rev. Toxicol. 46 (10), 835–844.
Sonich-Mullin, C., Fielder, R., Wiltse, J., Baetcke, K., Dempsey, J., Fenner-Crisp, P., et al., 2001. IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis. Regul. Toxicol. Pharmacol. 34 (2), 146–152.
Strickland, J., Clippinger, A.J., Brown, J., Allen, D., Jacobs, A., Matheson, J., et al., 2018. Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies. Regul. Toxicol. Pharmacol. 94, 183–196.
Strickland, J., Daniel, A.B., Allen, D., Aguila, C., Ahir, S., Bancos, S., et al., 2019. Skin sensitization testing needs and data uses by US regulatory and research agencies. Arch. Toxicol. 93 (2), 273–291.
Suntharalingam, G., Perry, M.R., Ward, S., Brett, S.J., Castello-Cortes, A., Brunner, M.D., et al., 2006. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. N. Engl. J. Med. 355 (10), 1018–1028.
Swenberg, J.A., Lehman-McKeeman, L.D., 1999. Alpha 2-Urinary globulin-associated nephropathy as a mechanism of renal tubule cell carcinogenesis in male rats. IARC Sci. Publ. 147, 95–118.
Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., et al., 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul. Toxicol. Pharmacol. 71 (2), 337–351.
Wang, J., Li, Z., Sun, F., Tang, S., Zhang, S., Lv, P., et al., 2017. Evaluation of dermal irritation and skin sensitization due to vitacoxib. Toxicol Rep 4, 287–290.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018.