Toxicology in Vitro 46 (2018) 163–165

Contents lists available at ScienceDirect

Toxicology in Vitro

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

Validation redefined

Aldert H. Piersma[a,b,*], Jan van Benthem[a], Janine Ezendam[a], Anne S. Kienhuis[a]

[a] Center for Health Protection, National Institute for Public Health and the Environment RIVM, The Netherlands
[b] Institute for Risk Assessment Sciences IRAS, Utrecht University, The Netherlands

This paper discusses needs and opportunities to redefine validation of alternative methods to animal testing. *In vitro* chemical hazard assessment is moving from individual assays to combinations of assays in batteries and testing strategies guided by adverse outcome pathways. This has consequences for the way individual assays and testing strategies are validated. We propose that quality criteria of reproducibility and transferability and description of chemical applicability domain remain essential at the level of individual assays. However, validation in terms of predictivity of individual assays based on a variety of chemicals is no more relevant. Rather, sufficient coverage of the biological domain studied by a battery of complementary assays should be the prime determinant of the validity of test batteries and testing strategies.

Traditionally, chemical hazard and risk assessment for man is assessed by means of experimental animal studies. Global guidelines for animal test methods were established in the early nineteen eighties. Soon thereafter the call arose for alternatives that would Refine, Reduce and/or Replace (3R) animal use, in line with the principles established in the fifties by Russell and Burch (1959). A host of *in vitro* assays appeared, originating either from existing biological research models that were transformed to toxicity assays, or from novel assays based on primary or continuous cell cultures up to tissue and organ cultures. Characteristics were defined that would be required for an animal-free assay to be applicable for chemical hazard assessment (Hartung et al., 2004). A definition of the biological domain covered by the assay is needed to determine its applicability in terms of modes of action covered. The reproducibility of test results within and among laboratories are considered important aspects of robustness and a prerequisite for general applicability of alternative tests. Furthermore, the chemical applicability domain is important in that it describes the limitations of the assays in terms of chemicals that cannot be tested due to *e.g.* solubility or volatility issues. Another crucial aspect relates to the predictive capacity of animal-free assays. Validation of an assay basically entails the evaluation of the combination of these characteristics. Practical approaches towards validation were formulated by the European Center for Validation of Alternative Methods (Clothier and Balls, 1990; Balls, 1995). Strictness *versus* flexibility of validation rules have been important discussion items. These approaches have been implemented in an OECD Guidance Document No. 34 on validation of test methods (OECD, 2005).

OECD GD No. 34 defined validation as 'the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose'. Initially, animal-free assays were validated for the one-in-one replacement of existing animal tests. This required that the predictive capacity of these tests for the *in vivo* situation should be established. Therefore, validation studies were designed with the aim to compare outcomes of individual assays with those of the same chemicals tested in animal studies. The result obtained in animal-free assays was often characterized in a qualitative sense as positive *versus* negative. Results of a series of reference chemicals were combined to calculate overall test sensitivity, specificity and predictive capacity based on qualitative *in vitro-in vivo* comparisons. The number of chemicals tested, and with that the chemical space covered, in any such validation study was pragmatically limited. Validation studies ended up around 80% predictivity in many such validation studies, *e.g.* for developmental toxicity (Brown, 1987; Genschow et al., 2002) and skin sensitization (Bauch et al., 2011; Natsch et al., 2011), which was usually considered a good result. Based on such validation outcomes, combined with complying with the requirements mentioned above, the alternative assay was considered validated and valid for use as a replacement of the animal test that it was intended to replace.

Three decades of expertise with this validation process have increasingly revealed its limitations. First, taking the animal study as the gold standard for comparison of novel assays appeared flawed, as the species of interest in human health risk assessment is man. Animal studies do not always correctly predict toxicity in man, be it the nature, severity, or the dosimetry of the adverse health effect involved. Moreover, the presumed one-in-one replacement is overly simplistic, as it does not take account of the reductionist nature of the *in vitro* test as compared to the whole organism animal test that it intends to replace. Furthermore, validations based on a necessarily limited number of chemicals do not necessarily predict test performance for yet another set of chemicals (Paquette et al., 2008; Marx-Stoelting et al., 2009). Predictivity percentages from such validation studies therefore have little significance for estimating general test performance. Finally, the qualitative nature of many *in vitro* test outcomes (positive/negative) leaves no room for quantitative assessment and for extrapolation to expected effective doses in the *in vivo* animal study. In the past decades,

* Corresponding author at: RIVM, Antonie van Leeuwenhoeklaan 9, P.O. Box 1, 3720 BA Bilthoven, The Netherlands.
E-mail address: aldert.piersma@rivm.nl (A.H. Piersma).

http://dx.doi.org/10.1016/j.tiv.2017.10.013
Received 15 June 2017; Received in revised form 3 October 2017; Accepted 8 October 2017
Available online 09 October 2017
0887-2333/

these limitations have probably played an important role in the reluctance to accept animal-free methods as replacements in regulatory frameworks for human chemical safety assessment (Adler et al., 2011; Piersma et al., 2014).

Whilst the limitations of classical validation of animal-free assays became apparent, toxicological hazard and risk assessment sciences changed gears significantly. The advent of the Tox21 initiative by the US National Academy of Sciences advocated founding human risk assessment on human data, collected either in the clinical setting or from animal-free assays incorporating human biological material (USA, 2007; Krewski et al., 2010). The large-scale ToxCast project of the US Environmental Protection Agency (EPA) started collecting and analyzing data of thousands of chemicals in hundreds of high throughput biological and biochemical assays to predict toxicity in the intact organism (Kavlock et al., 2012). OECD embraced the Adverse Outcome Pathway (AOP) concept. AOPs are used as a framework to structure all available information at different levels of biological complexity, *e.g.* molecular, cellular, organ and organism level, relevant to a particular adverse outcome (Ankley et al., 2010). AOPs have shown to be of value in the development of mechanism-based strategies that combined non-animal methods addressing different key events that lead to a specific adverse outcome. In the field of skin sensitization this has been quite successful as several AOP-based defined approaches are more accurate in predicting human skin sensitization hazard than the traditional animal method (Ezendam et al., 2016). In addition to the biological perspective of AOPs, the OECD QSAR toolbox provides a means of integration of chemical and toxicological information for the application of category approaches (Dimitrov et al., 2016). These models provide the basis for Integrated Approaches to Testing and Assessment (IATA) and Defined Approaches, which promote mechanism of action based hazard and risk assessment grounded on information of combinations of mechanistically relevant animal-free methods relevant for the specific toxicological question at stake (Tollefsen et al., 2014). A recent overview of relevant alternative technologies is given in (Eskes and Whelan, 2016).

These innovative ideas require a novel overarching approach of toxicological testing for regulatory purposes, which steps away from the animal as the gold standard, away from qualitative *in vitro* effect assessment as sufficient for hazard identification, and away from one-in-one replacement of animal studies. Thus, classical validation in terms of qualitative predictivity assessment in a one-in-one comparison with an animal study becomes obsolete. The question remains how the term 'validation' can be redefined to support the application of the novel IATA paradigm. Given the reductionist nature of animal-free assays, be they receptor binding assays, enzyme inhibition assays, cell proliferation or differentiation assays, or organs-on-a-chip, *etcetera*, paramount is the understanding of the biological domain that is incorporated in the system. For example, an estrogen antagonist will not show its antagonism in an assay that does not contain the estrogen receptor. Such an assay should not be used for the evaluation of estrogen antagonism. Thus, rather than statistical validation leading to predictivity percentages, mechanistic validation focused on carefully defining the biology of the system and its response characteristics is needed to be able to optimally use the assay in the context of a testing strategy. Continuing on the given example, the functionality of the estrogen test system can simply be shown by testing a limited number of agonists, antagonists, and negative controls. Validation with a wider chemical space is beyond what is needed to consider an individual test fit for purpose from the perspective of the biological domain covered. In addition to a definition of the biological system covered in the assay, what remains valid of the original validation paradigm is the continuing need to establish assay variability, reproducibility, and intra- and inter-laboratory transferability. These aspects are fundamental for understanding the reliability of any given biological assay. Likewise, the exposure situation in *in vitro* assays is fundamentally different from the *in vivo* situation. This may limit the applicability of *in vitro* assays as to certain chemical characteristics such as solubility and volatility. Therefore, defining the chemical applicability domain remains important to delineate limitations of *in vitro* assays.

Innovative hazard and risk assessment ideas often advocate a clean sheet case-by-case approach based on all non-testing information (*e.g.* chemical structure, physicochemical properties, structure-activity relationships, read across) available before any biological effect testing is initiated. This requires that for case-by-case recruitment a tool box of animal-free assays must be available, which contains all essential aspects of physiology that may need to be tested. The ToxCast initiative has made significant progress in this respect (Judson et al., 2015). However, the integration of information from all assays, and the extrapolation to predictions as to effects in the intact individual still meets with significant issues. The AOP paradigm defines key events in linear mono-directional cascades from molecular initiating event to adverse outcome, which may be monitored in dedicated *in vitro* assays. In the intact organism, most likely a multidimensional multidirectional network of closely interacting AOP's determines overall toxicity. This requires integration of AOP information at a higher level of abstraction for toxicity predictions to become realistic. It also requires understanding of the role of biokinetics in *in vitro* systems(Groothuis et al., 2015), and that *in vitro* concentration-response information is translated to *in vivo* effective dosages using quantitative extrapolation models, also referred to as reverse dosimetry (Yoon et al., 2012). Projects designing virtual liver and virtual embryo computational models have taken up the challenge of information integration into models of physiological processes (Kavlock, 2010; Leung et al., 2016; Hutson et al., 2017). Other initiatives are directed at defining ontologies in which (part of) the physiological system is systematically mapped, and can be modelled quantitatively in an all-encompassing computational toxicology system (Brinkley et al., 2013; Puelles et al., 2013). This computational system could then be fed with information from *in vitro* assays on the modulation by test compounds of selected key events, followed by integration of information at the level of the computational model, providing integrated predictions for toxicity in the intact organism, that is, in man.

Ultimately, the desirable prospect is that human hazard and risk assessment will be based on data from groups of dedicated *in vitro* assays and *in silico* models. Assays and models are selected and applied stepwise in battery and tiered approaches case-by-case, dependent on the test compound of interest and its effect pattern. These data are fed into the ontology-driven computational model, which will assess toxicity at the level of the intact body. Briefly, in this context, ontology is defined as a quantitative description of the integral network of adverse outcome pathways. The ontology allows identification of rate-limiting key events that need assessment in dedicated *in vitro* assays. The results of testing in this assay battery provide input for the computational model that predicts toxicity. This approach continues to require that individual assays are well characterized in terms of (a) their biological mechanism(s), (b) their reliability/reproducibility, and (c) their chemical applicability domain. Individual assays would be fit for purpose under the innovative paradigm if conforming to these three key validation characteristics of alternative assays. However, overall predictivity of individual assays for toxicity in the intact organism is not relevant in this approach. Rather, confidence in sufficient coverage of the biology by the combined battery of fit-for-purpose assays, reflecting the essential key events in the ontology, should suffice for considering a testing strategy valid. We do realize that confidence in this approach is not easily attained and can only be achieved by comprehensively demonstrating its performance. Case studies with selected chemicals using the integrated computational model and its panel of supporting *in vitro* assays can be employed to assess and optimize test strategy performance.

The basic philosophy underlying this approach is not that different to what lead to using animal studies in the past. Over half a century ago animal studies were introduced as test systems for human safety,

because the biology of the (animal) system was considered to provide the best possible model for predicting (human) hazard and risk. In the 21st century, given tremendous progress over the past half century in molecular, physiological, chemical and toxicological knowledge and tools, including dedicated human systems, the game change moving towards introducing *in vitro* and *in silico* modelling for human hazard and risk assessment is timely and feasible.

Transparency document

The http://dx.doi.org/10.1016/j.tiv.2017.10.013 associated with this article can be found, in online version.

References

Adler, S., Basketter, D., Creton, S., Pelkonen, O., van Benthem, J., Zuang, V., Andersen, K.E., Angers-Loustau, A., Aptula, A., Bal-Price, A., Benfenati, E., Bernauer, U., Bessems, J., Bois, F.Y., Boobis, A., Brandon, E., Bremer, S., Broschard, T., Casati, S., Coecke, S., Corvi, R., Cronin, M., Daston, G., Dekant, W., Felter, S., Grignard, E., Gundert-Remy, U., Heinonen, T., Kimber, I., Kleinjans, J., Komulainen, H., Kreiling, R., Kreysa, J., Leite, S.B., Loizou, G., Maxwell, G., Mazzatorta, P., Munn, S., Pfuhler, S., Phrakonkham, P., Piersma, A., Poth, A., Prieto, P., Repetto, G., Rogiers, V., Schoeters, G., Schwarz, M., Serafimova, R., Tahti, H., Testai, E., van Delft, J., van Loveren, H., Vinken, M., Worth, A., Zaldivar, J.M., 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. Arch. Toxicol. 85 (5), 367–485.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29 (3), 730–741.

Balls, M., 1995. Defining the role of ECVAM in the development, validation and acceptance of alternative tests and testing strategies. Toxicol. in Vitro 9 (6), 863–869.

Bauch, C., Kolle, S.N., Fabian, E., Pachel, C., Ramirez, T., Wiench, B., Wruck, C.J., Ravenzwaay, B.V., Landsiedel, R., 2011. Intralaboratory validation of four in vitro assays for the prediction of the skin sensitizing potential of chemicals. Toxicol. in Vitro 25 (6), 1162–1168.

Brinkley, J.F., Borromeo, C., Clarkson, M., Cox, T.C., Cunningham, M.J., Detwiler, L.T., Heike, C.L., Hochheiser, H., Mejino, J.L.V., Travillian, R.S., Shapiro, L.G., 2013. The ontology of craniofacial development and malformation for translational craniofacial research. Am. J. Med. Genet. C: Semin. Med. Genet. 163 (4), 232–245.

Brown, N.A., 1987. Teratogenicity testing in vitro: status of validation studies. Arch. Toxicol. Suppl. 11, 105–114.

Clothier, R.H., Balls, M., 1990. Validation of alternative toxicity tests: Principles, practices and cases. Toxicol. in Vitro 4 (4), 692–693.

Dimitrov, S.D., D., R., Sobanski, T., Pavlov, T.S., Chankov, G.V., Chapkanov, A.S., Karakolev, Y.H., Temelkov, S.G., Vasilev, R.A., Gerova, K.D., Kuseva, C.D., Todorova, N.D., Mehmed, A.M., Rasenberg, M., Mekenyan, O.G., 2016. QSAR Toolbox – workflow and major functionalities. SAR QSAR Environ. Res. 27 (3), 203–219.

Eskes, C., Whelan, M., 2016. Validation of Alternative Methods for Toxicity Testing. Springer Verlag.

Ezendam, J., Braakhuis, H.M., Vandebriel, R.J., 2016. State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. Arch. Toxicol. 90 (12), 2861–2883.

Genschow, E.S.H., Scholz, G., Seiler, A., Brown, N., Piersma, A., Brady, M., Clemann, N., Huuskonen, H., Paillard, F., Bremer, S., Becker, K., 2002. The ECVAM international validation study on in vitro embryotoxicity tests: results of the definitive phase and evaluation of prediction models. European Centre for the Validation of Alternative Methods. Altern. Lab. Anim 30 (2), 151–176.

Groothuis, F.A., Heringa, M.B., Nicol, B., Hermens, J.L.M., Blaauboer, B.J., Kramer, N.I., 2015. Dose metric considerations in in vitro assays to improve quantitative in vitro–in vivo dose extrapolations. Toxicology 332 (Supplement C), 30–40.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A.J., Prieto, P., Sabbioni, E., Scott, L., Worth, A., Zuang, V., 2004. A modular approach to the ECVAM principles on test validity. Altern. Lab. Anim 32 (5), 467–472.

Hutson, M.S., Leung, M.C.K., Baker, N.C., Spencer, R.M., Knudsen, T.B., 2017. Computational Model of Secondary Palate Fusion and Disruption. Chem. Res. Toxicol. 30 (4), 965–979 (Apr 17).

Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D.M., Filer, D.L., Houck, K.A., Martin, M.T., Sipes, N., Richard, A.M., Mansouri, K., Setzer, R.W., Knudsen, T.B., Crofton, K.M., Thomas, R.S., 2015. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. Toxicol. Sci. 148 (1), 137–154.

Kavlock, R.D., 2010. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J. Toxicol. Environ. Health B Crit. Rev. 13 (2–4), 197–217.

Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., Richard, A., Rotroff, D., Sipes, N., Dix, D., 2012. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem. Res. Toxicol. 25 (7), 1287–1302.

Krewski, D., Acosta, D., Andersen, M., Anderson, H., Bailar, J.C., Boekelheide, K., Brent, R., Charnley, G., Cheung, V.G., Green, S., Kelsey, K.T., Kerkvliet, N.I., Li, A.A., McCray, L., Meyer, O., Patterson, R.D., Pennie, W., Scala, R.A., Solomon, G.M., Stephens, M., Yager, J., Zeise, L., 2010. Toxicity testing in the 21st century: a vision and a strategy. J. Toxicol. Environ. Health B Crit. Rev. 13 (0), 51–138.

Leung, M.C.K., Hutson, M.S., Seifert, A.W., Spencer, R.M., Knudsen, T.B., 2016. Computational modeling and simulation of genital tubercle development. Reprod. Toxicol. 64, 151–161.

Marx-Stoelting, P., Adriaens, E., Ahr, H.J., Bremer, S., Garthoff, B., Gelbke, H.P., Piersma, A., Pellizzer, C., Reuter, U., Rogiers, V., Schenk, B., Schwengberg, S., Seiler, A., Spielmann, H., Steemans, M., Stedman, D.B., Vanparys, P., Vericat, J.A., Verwei, M., van der Water, F., Weimer, M., Schwarz, M., 2009. A review of the implementation of the embryonic stem cell test (EST). The report and recommendations of an ECVAM/ReProTect Workshop. Altern. Lab. Anim 37 (3), 313–328.

Natsch, A., Bauch, C., Foertsch, L., Gerberick, F., Norman, K., Hilberer, A., Inglis, H., Landsiedel, R., Onken, S., Reuter, H., Schepky, A., Emter, R., 2011. The intra- and inter-laboratory reproducibility and predictivity of the KeratinoSens assay to predict skin sensitizers in vitro: results of a ring-study in five laboratories. Toxicol. in Vitro 25, 733–744.

OECD, 2005. Guidance Document No. 34 on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. OECD Environment, Health and Safety Publications Series on Testing and Assessment.

Paquette, J.A., Kumpf, S.W., Streck, R.D., Thomson, J.J., Chapin, R.E., Stedman, D.B., 2008. Assessment of the embryonic stem cell test and application and use in the pharmaceutical industry. Birth Defects Res. B Dev. Reprod. Toxicol. 83 (2), 104–111.

Piersma, A.H., Ezendam, J., Luijten, M., Muller, J.J.A., Rorije, E., van der Ven, L.T.M., van Benthem, J., 2014. A critical appraisal of the process of regulatory implementation of novel in vivo and in vitro methods for chemical hazard and risk assessment. Crit. Rev. Toxicol. 24 (0), 1–19.

Puelles, L., Harrison, M., Paxinos, G., Watson, C., 2013. A developmental ontology for the mammalian brain based on the prosomeric model. Trends Neurosci. 36 (10), 570–578.

Russell, W.M.S., Burch, R.L., 1959. The Principles of Humane Experimental Technique., Methuen, London. Reprinted by UFAW, 1992: 8 Hamilton Close, South Mimms, Potters Bar, Herts EN6 3QD England (1959).

Tollefsen, K.E., Scholz, S., Cronin, M.T., Edwards, S.W., de Knecht, J., Crofton, K., Garcia-Reyero, N., Hartung, T., Worth, A., Patlewicz, G., 2014. Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul. Toxicol. Pharmacol. 70 (3), 629–640.

USA, N. N. A. o. S, 2007. Toxicity Testing in the Twenty-first Century: A Vision and a Strategy. National Academies Press, Washington, D.C.

Yoon, M., Campbell, J.L., Andersen, M.E., Clewell, H.J., 2012. Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results. Crit. Rev. Toxicol. 42 (8), 633–652.