National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

Development and implementation of alternative methods in reproductive toxicology

RIVM report 340700005/2011

A. H. Piersma

National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

Development and implementation of alternative methods in reproductive toxicology

RIVM Report 340700005/2011

Colophon

Aldert H. Piersma, RIVM/GBO

Contact:
Aldert H. Piersma
RIVM/GBO
aldert.piersma@rivm.nl

This investigation has been performed by order and for the account of the Ministry of Health, Welfare and Sports, within the framework of project V/340700: Expertise and Advice on Carcinogenicity, Mutagenicity and Reproductive Toxicity

Rapport in het kort

Ontwikkeling en invoering van alternatieve methoden in de reproductietoxicologie
Het RIVM probeert alternatieve testmethoden voor de zogeheten reproductietoxicologie zodanig te verbeteren dat ze betrouwbaar genoeg zijn om in de regelgeving te kunnen worden ingevoerd. Op die manier kan het aantal dierproeven worden verminderd. Reproductietoxicologie houdt zich bezig met mogelijke schadelijke effecten van stoffen op de vruchtbaarheid, de voortplanting, en de ontwikkeling van het ongeboren kind. In dit rapport heeft het RIVM, in opdracht van het ministerie van VWS, de stand van zaken beschreven van alternatieve testmethoden voor deze wetenschap.

Onderzoek naar alternatieve testmethoden
Het RIVM ontwikkelt zelf alternatieve tests en teststrategieën waarvan de wetenschap verwacht dat ze schade aan de ontwikkeling kunnen voorspellen. De afgelopen vijf jaar is voortgang geboekt op het gebied van de embryonale stamceltest, de ratten-embryokweek, en de zebravis-embryotest. Deze embryo's zijn niet levensvatbaar, waardoor deze organismen, evenals de stamcellen, volgens Europese wetgeving niet als proefdieren worden beschouwd. Bij deze tests wordt gekeken naar de effecten die een stof op het niveau van genen veroorzaakt (genexpressie). Effecten op dit niveau kunnen mogelijk subtieler voorspellen of en in welke mate stoffen schadelijk zijn. Bovendien is de verwachting dat effecten op genniveau in deze testen beter te vertalen zijn naar effecten voor de mens. De huidige wijze waarop effecten van stoffen worden vastgesteld, is gebaseerd op gezondheidsschade in proefdieren.

Modernisering regelgeving
Daarnaast is het RIVM actief in internationale verbanden om de regelgeving rond het gebruik van dierproeven en alternatieven te moderniseren. Het eigen onderzoek is een belangrijke ondersteuning hierbij.

Trefwoorden: Alternatieve testmethoden, reproductietoxicologie, teststrategie, embryokweek, embryonale stamceltest, zebravis embryotest, vermindering van dierproeven, genexpressie, innovatie

Abstract

Development and implementation of alternative methods in reproductive toxicology
RIVM is improving alternative test methods in the area of reproductive toxicology, to make them ready for regulatory application. This will reduce experimental animal use. Reproductive toxicology pertains to adverse effects on fertility and reproduction and on prenatal development of the unborn child. In this report, the state of the art of alternatives in this area is reviewed.

Research into alternative methods
RIVM develops alternative tests and testing strategies expected to be able to predict reproductive toxicity. During the last five years advances have been achieved with the embryonic stem cell test, the rat embryo culture and the zebra fish embryo test. These embryos are highly premature and not considered as experimental animals under European law. In these tests effects are studied on the level of gene expression. Effects on gene expression may give more detailed information on compound effects. Moreover, such effects may allow easier translation to effects in man. Current effect assessment is done in experimental animals.

Innovation of regulatory toxicology
In addition, RIVM is active in the international arena to innovate regulations concerning experimental animals and alternative test systems. This research provides important support for these activities.

Keywords: alternative test methods, reproductive toxicology, testing strategies, embryo culture, embryonic stem cell test, zebra fish embryo test, reduction experimental animal use, gene expression, innovation

Contents

Summary—6

1 Introduction—7

2 Rat post-implantation whole embryo culture—9

3 Embryonic stem cell test—10

4 Zebra fish embryo test—12

5 Testing strategy issues—13

6 Relevant publications in 2010 and 2011—16

Appendix—19

Summary

This report gives an overview of recent developments in the area of alternatives to animal experimentation in reproductive toxicology, in particular the research carried out under the "*kennisvraag advisering en onderzoek carcinogenese, mutagenese en reproductietoxicologie*" at the order of the Dutch Ministry for Public health, Welfare and Sport. The project aims among other things, at advancing the development and implementation of alternative methods in reproductive toxicology. The project participates actively in the development of rat whole embryo culture, mouse embryonic stem cell test and zebra fish embryo test as alternative assays that can reduce animal use in regulatory toxicology. The added value of differential gene expression assessment in detecting and predicting developmental toxicity in these assays is extensively being studied. Finally, knowledge built up within the project is combined with existing knowledge to contemplate innovations in the hazard assessment testing strategy for developmental toxicity. International embedding of this discussion is aimed for to facilitate reduction of animal use and enhancement of mechanistic hazard assessment, improving the safety assessment of chemicals.

1 Introduction

This report gives an overview of recent developments in the area of alternatives to animal experimentation in reproductive toxicology, in particular the research carried out under the "*kennisvraag advisering en onderzoek carcinogenese, mutagenese en reproductietoxicologie*" at the order of the Dutch Ministry for Public Health, Welfare and Sport. The project aims among other things, at advancing the development and implementation of alternative methods in reproductive toxicology. This field is characterised by relatively high animal use in regulatory toxicology, using around 60% of all experimental animals under the REACH legislation. The necessity for advancing alternative methods is likewise high and efforts have been underway for three decades to design useful alternatives. Their implementation has lagged behind, partly because of uncertainties about their applicability domain and their predictive value. A complicating factor is that the reproductive cycle is a complex system, combining many mechanisms of development and toxicity that cannot readily be mimicked *in vitro* in a comprehensive way. Therefore, efforts are underway to better define alternative tests as to the mechanism of action they include, and to do comparative studies among alternative assays and with animal test results to provide a framework for translating results from alternative assays to the situation in the intact animal. In addition, the combination of complementary assays in a testing strategy is envisaged, which might result in enhanced predictive power as compared to individual assays. RIVM has particularly focused on three alternatives that currently enjoy major developments in research globally. In the following, these assays and the work done on them by RIVM in the past two years are summarised. Finally, for further information, the greater context of alternative testing batteries and the implementation of alternative assays in regulatory toxicology is comprehensively addressed in an appendix text, which has been accepted as a book chapter in a toxicological series.

The rat whole embryo culture assay is based on the culture of rat embryos between gestation days 10 and 12. During this period of embryogenesis a major part of the essential morphogenetic process of organogenesis takes place. The heart is formed and starts beating, the neural tube closes and forms the various brain compartments and the spinal cord. The segmented structure of the vertebral column is formed, craniofacial structures such as eyes, ears, olfactory bulbs, branchial bars, maxillary and mandibular processes develop, as do the first stages of limb bud development. The development of each of these structures can be monitored in culture. During this stage the embryo is still independent of the placenta and develops almost at the same rate as *in vivo*. Development of each of the organ anlagen can be morphologically scored and allows the assessment of the effects of chemicals. The embryo developing separately from the mother offers advantages, as only direct effects are studied and not those secondary to maternal toxicity. In addition, the compound concentration in the culture medium is relevant as comparative to the dose at target in *in vivo* embryotoxicity studies. We have used this method for over twenty years now as an alternative model for embryotoxicity testing. In the current project we are employing differential gene expression assessment as a tool to study embryotoxicity in a more detailed and mechanistic way. The aim is to gain knowledge on applicability domain and predictivity with a view to enhancing the implementation of the assay as an alternative to full *in vivo*

animal testing in developmental toxicology. This research is described in more detail in chapter 2.

The embryonic stem cell test (EST) is based on mouse pluripotent embryonic stem cells. These cells share many characteristics with the inner cell mass of the blastocyst in very early embryonic development. They can differentiate into all the tissues of the animal, dependent on the culture conditions applied. These cells offer the possibility to study effects of chemicals on differentiation processes that are established in *in vitro* models. The EST has been studied since the late 1990s, and was standardised as an ECVAM method. It is a genuine alternative assay as it is entirely animal-free. It is employed widely in research departments throughout the world, often as a pre-screening assay to collect preliminary information on effects that may give indications for developmental toxicity potential of the chemical under study. We have introduced the cardiac muscle cell differentiation system over five years ago, and have studied the effects of chemicals on the differentiation process in various projects. Fundamental questions about this assay regard its applicability domain and its predictive value. It is not straightforward to extrapolate an effect on cardiac muscle cell differentiation to embryotoxicity *in vivo*. Likewise, predictivity estimated from a diverse series of chemicals tested may not be relevant for additional chemicals. We have applied category approaches to try and circumvent this problem. This showed us that within classes of chemicals, such as glycolethers and triazoles, the potency ranking of chemicals in EST was reminiscent of their *in vivo* potency ranking, enabling conclusions about the applicability of the EST for these particular chemical classes. As for WEC, in recent years we have applied differential gene expression as a tool to study embryotoxicity in a more detailed and mechanistic way. In chapter 3 we describe recent progress with this approach.

The zebra fish embryotoxicity test (ZET) employs postfertilisation embryos that are followed for 72 hours. During this time they will develop into complete, hatched and swimming larvae. This represents a very rapid development as compared to mammalian species and allows the study of chemical effects on the entire embryogenesis period. The body plan of vertebrates being highly conserved especially during organogenesis, this model is extensively used in developmental biology. Likewise, it is a useful model to study toxic effects of chemicals on embryogenesis. The eggs being transparent, morphogenesis can be followed continuously. Their natural environment being water, their culture is very easy. During the first 120 hours of development they are dependent on yolk sac feeding, and are therefore not considered experimental animals under European law. Therefore, this test can technically be regarded as animal-free. The test has become popular among reproductive toxicologists in the last decade and is increasingly being studied for its pros and cons as an alternative test for embryotoxicity. As for WEC and EST, in recent years we have applied differential gene expression as a tool to study embryotoxicity in a more detailed and mechanistic way. In chapter 4 we will address progress made with this alternative test.

2 Rat post-implantation whole embryo culture

This assay, established back in the 1970s, has survived four decades and is still used in various laboratories. The main characteristic of the assay is that it studies the development of the entire rat embryo in a crucial stage of organogenesis in isolation from the maternal animal in a culture flask. Test compounds can be added directly and malformations can be readily scored. Disadvantages include the need to isolate embryos from pregnant dams; therefore the test is not animal-free. We use it for bridging the gap between animal-free alternatives and the intact pregnant animal. We hypothesised that if gene expression changes induced by compound exposure should be similar between different test systems, this would provide compelling argumentation that the gene expression effects in the alternative assay would be relevant as a measure for toxicity induction in the whole animal. Therefore, this approach could provide a mechanistic validation of the method.

In our first study we investigated the differential gene expression response after exposure to retinoic acid. This chemical is the active form of vitamin A, which at physiological concentrations is an essential natural morphogen controlling vertebrate embryogenesis. At higher concentrations it becomes an embryotoxicant, which can give rise to severe malformations, as has been observed in children of mothers that took multivitamin preparations containing retinoic acid during pregnancy. Our study clearly showed retinoic acid specific gene expression changes in WEC, confirming that the model applying transcriptomics provides relevant gene expression responses and can potentially be used as a readout parameter for embryotoxicity assessment in the WEC.

In our second study we compared the gene expression signature of four diverse embryotoxicants, caffeine, methylmercury, the phthalate metabolite monobutylphthalate and the glycolether metabolite methoxyacetic acid. These compounds were tested for gene expression effects at concentrations that were similar in potency as regards morphological effects. We found very diverse gene expression signatures with each of these compounds in addition to differences in the abundance of the response. This study taught us that the same malformation pattern may be caused by very different molecular mechanisms. Furthermore, these data confirm that for predictions of embryotoxicity to be generated from gene expression data, it is necessary to combine responses of genes from different pathways in order to enhance the chances of detection. The challenge clearly is in the derivation of the best predictive gene set.

In our third study, we compared gene expression responses to methyl mercury in WEC with other *in vitro* and *in vivo* models, based on published literature data. There was a clear comparable response between those models mimicking some aspect of embryo development as opposed to models not incorporating developmental processes. This shows that the gene expression response that we observed in WEC is reminiscent of *in vivo* embryotoxicity, confirming that we are not measuring *in vitro* culture artefacts. This research was received by the scientific field as an important approach for convincing scientists and regulators about the applicability of alternative assays.

3 Embryonic stem cell test

This assay is at present perhaps the single most popular assay for research into alternatives for developmental toxicity testing. Many validation studies have shed light on the usefulness of the test, which showed a wide variety of predictability dependent on the actual chemicals tested. The classical readout of the test is the scoring under the microscope of beating heart muscle cell foci, which may be reduced through compound exposure. We figured that this readout was limited in its informative value and added differential gene expression assessment with the aim to enhance mechanistic knowledge on compound effects and to improve predictability as well as the extrapolation of effects to the intact animal and man.

Initial studies have addressed the question of how gene expression changes during progression of the differentiation process during the ten-day culture period in the EST. We established that the most dramatic changes in gene expression occurred between days three and four in the protocol, which represents the onset of the major differentiation in the assay. Moreover, variations in gene expression between replicates were smallest at that stage. Both aspects argue for this time point as the optimal one to study gene expression in the EST.

Follow-up studies have focused on studying gene expression effects of a variety of chemicals, leading to a present database of around 20 compounds. On the basis of this work, we defined a gene set of 52 genes that would be predictive in retrospect for all compounds studied except for two. One of them showed non-significant changes in gene expression only due to low dose exposure, the other showed very large variations in response with unknown cause, both of which are well-known reasons for failure of effect finding. In general terms these results were very promising and are employed for further analysis and comparison of gene expression signatures.

Other studies have addressed concentration-response relationships at the gene expression level. These are considered essential in view of the question as to what should be considered an adverse effect in an *in vitro* test situation. Usually, the morphological readout, inhibition of cardiac muscle cell differentiation in this case, is taken as a corollary of the adverse effect *in vivo*. However, this is debatable, as differentiation inhibition may or may not lead to adverse effects in the *in vivo* situation. Moreover, gene expression changes, occurring at lower concentrations than cellular differentiation inhibition, are even more difficult as to the interpretation of adversity. *In vivo* exposures will always lead to a physiological response, likely involving differential gene expression. It is only beyond the threshold of adversity that toxic effects appear. In our concentration-response study with flusilazole, we showed that differentiation inhibition and cytotoxicity, occurring with different concentration response characteristics, were nicely accompanied by similar concentration response curves at the level of related gene sets for development and cell cycle, respectively. This analysis gave us important clues about the interpretation of gene expression as compared to classic morphology.

Moreover, potency ranking can be done on the basis of gene expression studies. A study with four phthalate metabolites of different embryotoxic potency showed that EST could discriminate these compounds in terms of potency and moreover, could rank them in the same order of embryotoxic potency as was known from *in vivo* animal study data. This work is internationally renowned and receives much attention at scientific conferences. It also feeds in to discussions in international regulatory groups that contemplate the introduction of alternative methods in regulatory toxicology.

4 Zebra fish embryo test

This assay was introduced in developmental toxicity testing at RIVM two years ago. It is widely considered as a useful intermediate between cell culture alternatives on the one hand and mammalian animal tests on the other. The zebra fish embryo develops in five days from a fertilised egg into an almost fully developed hatched embryo. Its development can be followed easily through its transparent egg and chemicals can be added to the water medium for assessing toxic effects. Legally, at these early stages the embryo is not considered an experimental animal. We are employing the zebra fish embryo to perform similar gene expression studies as in the above models and to do comparative studies that should indicate the advantages and disadvantages of each of the systems.

We have defined a standardised system for scoring of embryonic development of the zebra fish in the first 72 hours after fertilisation. During this period, the entire body plan of the larva is formed, ending in hatching of a free swimming larva. A scoring system was developed which allows staging of embryonic development. Using a series of chemicals we showed that developmental delays induced by compound exposure could be monitored efficiently using the scoring system, which we published as the General Morphology Score (GMS).

Subsequent studies have addressed the usefulness of differential gene expression assessment by transcriptomics for detecting embryotoxicity. Triazole antifungals and glycolethers were tested and showed compounds class-specific gene expression signatures. The numbers of genes regulated after 24 hours exposure starting after fertilisation was roughly 50-fold higher after glycolether exposure as compared to triazole exposure. This was remarkable as equipotent concentrations were applied based on the GMS at 72 hours. However, examining the morphology at 24 hours, we confirmed that the glycolether exposed embryos were already malformed in contrast to those exposed to triazoles, which were morphologically normal. This showed that the pace of maldevelopment over time differs between compound classes. In addition, we defined class-specific gene expression signatures. These signatures will be instrumental in clarifying mechanisms of action and interpretation of embryotoxicity in view of extrapolation between species. The latter is essential for defining the use of alternative assays, as they are designed to predict human hazard and risk.

The effects of dose level are now investigated, in order to compare the relative sensitivity of morphologically observed effects *versus* gene expression changes. This research should enable us to better define those genes that show expression characteristics that correlate best with adverse effects observed. The study results show clear concentration-response characteristics on the level of individual genes as well as sets of functionally related genes. Gene sets and physiological pathways known to be affected by the test compound, such as embryonic development, retinoic acid metabolism, steroidogenesis and fatty acid metabolism, were clearly regulated. These findings confirm the specificity of the gene expression readout and provide a further basis for dose selection in this type of assay.

In the international scientific arena, the zebra fish assay is also considered as a useful alternative and the database underpinning its use is rapidly growing. RIVM contributes to these developments, which gives us an important leading role in test guideline and strategy development for implementation in regulatory toxicology in OECD and EU frameworks.

5 Testing strategy issues

What is the current status of development and implementation of alternative assays in reproductive toxicology? In the foregoing chapters we have shown that significant scientific progress has been made in three assays in which RIVM has invested under sponsorship of the VWS project. This work is part of worldwide research into assay development using transcriptomics as well as additional molecular technologies with which biological responses to toxicant exposures can be studied in great detail. Deciphering mechanisms of action is widely thought to provide important instruments for extrapolating *in vitro* findings to human hazard and risk. Our work is highly embedded in collaborations within the Netherlands Toxicogenomics Centre as well as with the USEPA National Center for Toxicological Research ToxCast program and the European Union ChemScreen project. This enables us as a government agency to play a significant role in the interface between scientific developments and the regulatory implementation of newly developed alternative test systems.

Why do we need to study three different alternative tests for developmental toxicity? A crucial issue with the reproductive cycle, including fertility, pregnancy, embryofoetal development, birth and postnatal development until adulthood, is its complexity and multifaceted nature. Not one single assay will be able to reflect all possible mechanisms of action and relevant reproductive processes. Therefore, in this area it is envisaged that a battery of complementary test systems will be necessary to cover all aspects of the reproductive cycle. We have focused on three assays for developmental toxicity, representing one important segment of the reproductive cycle, embryogenesis. These will have to be evaluated for predictivity and applicability domain before their optimal place in a testing strategy can be defined.

Why have so many years of research into alternatives resulted in so few implemented tests? The answer to this question relates largely to predictivity and applicability domain. Classical validation studies have focused on testing a series of unrelated chemicals and defining percentage of correct prediction in a binary fashion, e.g., using a +/– score. We have subsequently learned that this is a superficial assessment of the usefulness of a test. Additional sets of compounds showed very different predictiveness and mathematical prediction models failed dramatically. This has taught us that we should rather focus attention on mechanistic evaluation of the test. This means that knowledge of the mechanism(s) and end point(s) incorporated in the test should form the basis of assessing predictivity. Thus, an embryotoxicant which is negative in a given test may nevertheless have provided a useful readout if its mechanism of action is not present in the test. Combining assays with complementary biological characteristics would then lead to improved predictivity.

How are we going to promote implementation of alternatives in regulatory toxicology? The answer is in taking the consequence of the above discussion. We need to better define the applicability domains of *in vitro* assays and combine them in an integrated testing strategy, making sure that the most relevant biological pathways underlying reproductive toxicity are incorporated in the strategy. Initiatives towards this approach can be found in the OECD conceptual framework guidance document, in the ReProTect ring trial and currently, in the ChemScreen European Project. These studies are explained in greater detail in the appendix to this report. The common main line of these initiatives is to combine a series of complementary alternative assays into a testing strategy that should be applied before *in vivo* animal testing is done.

Moreover, the results of the *in vitro* battery of tests should inform about the necessity of subsequent *in vivo* testing. This could result in refraining from *in vivo* testing or fine-tuning *in vivo* testing on the basis of specific *in vitro* test findings. Whilst these strategies are being developed, their mechanistic basis should provide convincing argumentation to persuade the scientific and regulatory arenas of the acceptability of such an approach.

How do regulatory agencies and policy makers respond to these developments? There appears a clear political trend towards the wish to reduce animal testing. Public awareness has been growing over the years, and the European parliament has agreed that animal testing for cosmetic ingredients is no longer acceptable in the very near future. On the other hand, the REACH legislation has resulted in an increase in animal testing in order to update and complete toxicology dossiers of around 30,000 chemicals. The estimated extra animal cost may amount to 10 million experimental animals. This has all greatly stimulated research into alternative assays for toxicity testing. On the other hand, neither the scientific world nor the regulatory agencies seem to be ready for the rapid implementation of alternatives. A recent scientific assessment of the progress towards the implementation of alternatives for cosmetics ingredients testing was carried out under the leadership of ECVAM. This study clearly showed that in most if not all areas of toxicity testing alternatives cannot yet match the level of hazard identification currently gained from animal studies. Moreover, proposed innovations in testing strategies in REACH meet with significant opposition from ECHA, the European Chemicals Agency which is responsible for carrying out the REACH regulation. A recently accepted OECD guideline for an extended one-generation study (OECD TG 443) could potentially reduce animal use by 40% as compared to the currently used 2-generation study, and by no less than 15% in the entire REACH programme. This could amount to a reduction of several million experimental animals. In this particular example, although published scientific studies from an international group of experts have provided justification and advocated the changeover, ECHA experts in particular have opposed this change, based primarily on legal arguments. The reader should realise that this example concerns replacing an animal study for another animal study, whereas replacement of animal studies by alternatives is even more difficult. This illustrates that scientific progress is not the only a bottleneck for the implementation of alternatives, but that active interaction with policy makers and regulatory agencies is necessary to pave the way for change.

How has this work contributed to the aims of the VWS-sponsored RIVM project? First of all, by contributing to scientific progress in the development of alternative assays, RIVM plays an important role in directing scientific research into assay development, characterisation and evaluation in a way that facilitates implementation in regulatory toxicology. RIVM as a government agency is more equipped to take this perspective than universities or private institutes. Secondly, by contributing to the science, RIVM has gained a respected position in international discussion arenas on the subject. Knowledge equals expertise. Regulatory scientists can be informed directly from our laboratory experience. Furthermore, understanding the regulatory sensitivities, we have been able to contribute this perspective in various national and international projects that we are involved in, which invest in the design of innovative testing strategies in which alternative assays play key roles. Thus, RIVM continues to be a crucial and critical partner providing basic expertise-driven knowledge focused on the implementation of alternative assays in regulatory toxicology.

The philosophy behind these complex interactions and regulatory implications are further elaborated in the appendix to this report. The work on these various aspects is expected to be further extended in the coming years, as palatable spin-off in regulatory decision making appears to come closer. Serious attention in the area of cosmetics and in REACH for alternatives is emerging and our further efforts to support these developments with scientific expertise and knowledge provides an invaluable basis for further work towards the implementation of alternative methods in regulatory toxicology.

6 Relevant publications in 2010 and 2011

Piersma AH, van Dartel DAM, van der Ven LPM. Alternative Methods in Reproductive Toxicology. Book chapter, Comparative Toxicology 2nd Ed., Elsevier, Ed. Charlene McQueen; Chapter, Chapter 12.20 p 293-305, 2010.

Dorien A.M. van Dartel, Jeroen L.A. Pennings, Frederik J. van Schooten, Aldert H. Piersma. Transcriptomics-based Identification of Developmental Toxicants through their Interference with Cardiomyocyte Differentiation of Embryonic Stem Cells. Toxicology and Applied Pharmacology 243: 420–428, 2010.

Barbara Schenk, Marc Weimer, Susanne Bremer, Bart van der Burg, Rita Cortvrindt, Alexius Freyberger, Giovanna Lazzari, Cristian Pellizzer, Aldert Piersma, Wolfgang Schäfer, Andrea Seiler, Hilda Witters, Michael Schwarz. The ReProtect Feasibility Study, Reproductive Toxicology 30: 200-218, 2010.

Ahmed M. Osman, Dorien A.M. van Dartel, Edwin Zwart, Marco Blokland, Jeroen L.A. Pennings, Aldert H. Piersma. Proteome profiling of mouse embryonic stem cells to monitor cell differentiation and its modulation by monobutyl phthalate. Reproductive Toxicology 30 (2010) 322–332, 2010.

Theunissen PT, Schulpen SHW, van Dartel DAM, Hermsen SAB, van Schooten FJ, Piersma AH. An abbreviated protocol for multilineage neural differentiation of murine embryonic stem cells and its perturbation by methylmercury. Reproductive Toxicology 2010 Jul;29(4):383-92.

Mirjam Luijten, Vincent van Beelen, Aart Verhoef, Marc Renkens, Marcel van Herwijnen, Anja Westerman, Frederik-Jan van Schooten, Jeroen Pennings, Aldert H. Piersma. Transcriptomics analysis of retinoic acid embryotoxicity in rat post-implantation whole embryo culture. Reproductive Toxicology 30, 333-340, 2010.

Esther de Jong, Wout Slob, Aldert H. Piersma. Application of the benchmark approach in the correlation of *in vitro and in vivo* data in developmental toxicity. In: Methods in Bioengineering: Alternative Technologies to Animal Testing, ed. Tim Maguire & Eric Novik. Artech House, Boston, London, Chapter 9, pp. 159-169, 2010.

Dorien A.M. van Dartel, Jeroen L.A. Pennings, Liset J.J. de la Fonteyne, Marcel H. van Herwijnen, Joost H. van Delft, Frederik J. van Schooten, Aldert H. Piersma, Monitoring Developmental Toxicity in the Embryonic Stem Cell Test Using Differential Gene Expression of Differentiation-related Genes. Toxicol Sci. 2010 Jul;116(1):130-9.

Jochem Louisse, Esther de Jong, Johannes J.M. van de Sandt, Bas J. Blaauboer, Ruud A. Woutersen, Aldert H. Piersma, Ivonne M.C.M. Rietjens, Miriam Verwei. The use of *in vitro* toxicity data and physiologically based kinetic modelling to predict dose-response curves for *in vivo* developmental toxicity of glycol ethers in rat and man. Toxicol Sci. 118(2):470-484, 2010.

George P. Daston, Robert E Chapin, Anthony R. Scialli, Aldert H. Piersma, Edward W. Carney, John M. Rogers, Jan M. Friedman. A Different Approach to Validating Screening Assays for Developmental Toxicity. Birth Defects Research (Part B) 83:1–5, 2010.

Joshua F. Robinson, Vincent A. van Beelen, Aart Verhoef, Marc F.J. Renkens, Mirjam Luijten, Marcel H.M. van Herwijnen, Anja Westerman, Jeroen L.A. Pennings, Aldert H. Piersma. Embryotoxicant-specific transcriptomics responses in rat post-implantation whole embryo culture. Toxicological Sciences 118(2):675-685, 2010.

Dorien A.M. van Dartel, Jeroen L.A. Pennings, Liset J.J. de la Fonteyne, Karen J.J. Brauers, Sandra Claessen, Joost H. van Delft, Jos C.S. Kleinjans, Aldert H. Piersma. Evaluation of Developmental Toxicant Identification Using Gene Expression Profiling in Embryonic Stem Cell Differentiation Cultures. Toxicol Sci. 119(1):126-34. 2011.

Adler S, Basketter D, Creton S, Pelkonen O, van Benthem J, Zuang V, Andersen KE, Angers-Loustau A, Aptula A, Bal-Price A, Benfenati E, Bernauer U, Bessems J, Bois FY, 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 SB, 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, Tähti H, Testai E, van Delft J, van Loveren H, Vinken M, Worth A, Zaldivar JM. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. Arch Toxicol. 2011 May;85(5):367-485.

Pennings JL, van Dartel DA, Robinson JF, Pronk TE, Piersma AH. Gene set assembly for quantitative prediction of developmental toxicity in the embryonic stem cell test. Toxicology 2011: 284(1-3):63-71.

Pennings J, van Dartel D, Pronk T, Hendriksen P, Piersma A. Identification by gene co-regulation mapping of novel genes involved in embryonic stem cell differentiation. Stem Cells Dev. 20(1): 115-126, 2011.

Dorien A.M. van Dartel, Jeroen L.A. Pennings, Liset J.J. de la Fonteyne, Karen J.J. Brauers, Sandra Claessen, Joost H. van Delft, Jos C.S. Kleinjans, Aldert H. Piersma. Concentration-dependent Gene Expression responses to Flusilazole in Embryonic Stem Cell Differentiation Cultures. Toxicology and Applied Pharmacology 251, 110–118, 2011.

E. de Jong, A. Doedée, M.A. Reis-Fernandes, H. Nau, A.H.Piersma. Potency ranking of valproic acid analogues as to inhibition of cardiac differentiation of embryonic stem cells in comparison to their *in vivo* embryotoxicity. Reprod. Toxicol. Reprod Toxicol. 2011 May;31(4):375-82.

Sanne A.B. Hermsen, Evert-Jan van den Brandhof, Leo T.M. van der Ven, Aldert H. Piersma. Relative embryotoxicity of two classes of chemicals in a modified zebra fish embryotoxicity test and comparison with their *in vivo* potencies. Toxicol. In vitro, 25(3):745-53, 2011.

Dorien A.M. van Dartel, Jeroen L.A. Pennings, Joshua F. Robinson, Jos C.S. Kleinjans, Aldert H. Piersma. Discriminating classes of developmental toxicants using gene expression profiling in the embryonic stem cell test. Toxicol Lett. 201(2):143-51, 2011.

Bart van der Burg, E. Dinant Kroese, Aldert H. Piersma. Towards a Pragmatic Alternative Testing Strategy for the Detection of Reproductive Toxicants. Reproductive Toxicology 31(4), 558-561, 2011.

Esther de Jong, Marta Barenys, Sanne AB. Hermsen, Aart Verhoef, Bernadette C. Ossendorp, Jos GM. Bessems, Aldert H. Piersma. Comparison of the mouse Embryonic Stem cell Test, the rat Whole Embryo Culture and the Zebra fish Embryotoxicity Test as alternative methods for developmental toxicity testing of six 1,2,4-triazoles. Toxicol Appl Pharmacol. 253(2):103-11, 2011.

Sanne A.B. Hermsen, Tessa Pronk, Evert-Jan van den Brandhof, Leo T.M. van der Ven, Aldert H. Piersma. Class specific gene expression changes in the zebra fish embryo after exposure to three glycol ether alkoxy acids and two 1,2,4-triazole antifungals. Reprod. Toxicol. 32(2):245-52, 2011.

Dorien A.M. van Dartel, Aldert H. Piersma. The embryonic stem cell test combined with toxicogenomics as an alternative testing model for the assessment of developmental toxicity. Review, Reproductive Toxicology, Reprod Toxicol. 32(2):235-44, 2011.

Theunissen PT, Pennings J, Robinson JF, Kleinjans JCS, Piersma AH. Time-response evaluation by transcriptomics of methylmercury effects on neural differentiation of murine embryonic stem cells. Toxicol Sci. 122(2):437-47, 2011.

Joshua F. Robinson, Theunissen PT, van Dartel DA, Pennings JL, Faustman EM, Piersma AH. Comparison of MeHg-induced toxicogenomic responses across in vivo and in vitro models used in developmental toxicology. Reprod Toxicol. 2011 Sep;32(2):180-8, 2011.

Appendix

Innovations in testing strategies in reproductive toxicology

Abstract

Toxicological hazard assessment currently finds itself at a crossroads where the existing classical test paradigm is challenged by a host of innovative approaches. Animal study protocols are being enhanced for additional parameters and improved for more efficient effect assessment with reduced animal numbers. Whilst existing testing paradigms have generally proven conservative for chemical safety assessment, novel alternative *in silico* and *in vitro* approaches and assays are being introduced that begin to elucidate molecular mechanisms of toxicity. Issues such as animal welfare, alternative assay validation, endocrine disruption and the US-NAS report on toxicity testing in the 21st century have provided directionality to these developments. The reductionistic nature of individual alternative assays requires that they be combined in a testing strategy in order to provide a complete picture of the toxicological profile of a compound. One of the challenges of this innovative approach is the combined interpretation of assay results in terms of toxicologically relevant effects. Computational toxicology aims at providing that integration. In order to progress, we need to follow three steps: 1) Learn from past experience in animal studies and human diseases about critical end points and pathways of toxicity. 2) Design alternative assays for essential mechanisms of toxicity. 3) Build an integrative testing strategy tailored to human hazard assessment using a battery of available alternative tests for critical end points that provides optimal *in silico* and *in vitro* filters to upgrade toxicological hazard assessment to the mechanistic level.

1. Chemical hazard and risk assessment

1.1 Tonnage-dependent risk assessment

The extent of animal testing in current chemical hazard and risk assessment in Europe, under the REACH legislation, depends on the annual tonnage level of production ***(1)***. The tonnage level is taken as a correlate of foreseen exposure and directly determines the level of detail of hazard assessment. For reproductive and developmental toxicity, no specific testing requirements exist at the base set level (<1 tonne per annum (tpa)). At this production level, usually only a 28-day sub-acute toxicity study (OECD407) is required in which effects on the male and female gonads can be indicative of reproductive toxicity ***(2)***. At higher tonnage levels, the OECD421 reproductive toxicity screening assay, which can be combined with a 28-day sub-acute toxicity study (OECD421/422), is the first indicator of possible reproductive effects in a functional reproductive test including mating, pregnancy and birth. However, this test, if showing no adverse results, cannot be taken as definitive proof of absence of reproductive toxic properties of the test compound. The ultimate tests required to fully address reproductive and developmental toxicity are the two-generation reproductive toxicity study (OECD416) and the prenatal

developmental toxicity study (OECD414). Each tonnage level has its own minimal requirements, which can be enhanced on a case by case basis if mandatory tests give reason for concern, warranting that further testing is not postponed to a higher tonnage level. In addition, the prenatal developmental toxicity study may be repeated in a second species, usually the rabbit after the rat, in case of equivocal results in the first species. The entire test package is then considered in order to determine the reproductive and overall no-observed-effect-level, which leads into risk assessment by determining regulatory exposure limits for man.

1.2 Classification and labelling

The globally harmonised system (GHS) for classification and labelling of hazardous substances is based on hazard, determined by the intrinsic properties of a substance to cause toxic effects ***(3)***. Hazard assessment for classification and labelling is based on the same animal study protocols that are required for the risk assessment of substances, as explained above. For substances toxic to fertility or to development, it is the specificity of the type of effect, in view of general toxicity that may occur at the same or higher dosages, that determines the classification. Rather than considering an integral risk assessment for the substance evaluated, the classification is based on hazard, “under reasonably expected use”. The latter could be limited to intended use, e.g., in an industrial setting or in household applications, but could also be interpreted to include accidental exposures. In Europe, a classified substance is subject to a host of downstream regulations, limiting its permitted applications in consumer products. The scientific justification of this system, basing risk management decisions on hazard assessment only, without complete risk assessment considering important aspects such as potency and actual human exposures, is subject to continued debate.

1.3 Innovations in animal protocols

The two-generation reproductive toxicity study (OECD416) and the prenatal developmental toxicity study (OECD414) have served us relatively well for three decades. No major calamities have been observed with registered compounds during that time. Nevertheless, novel aspects of reproductive toxicity and considerations of animal use and economy have prompted innovations that are currently subject to both experimental research as well as regulatory implementation.

Secondary to the societal concerns around chemical related endocrine disruption, the OECD407 sub-acute 28-day toxicity study protocol has been updated in 2007 with parameters relating to endocrine homeostasis. Specifically, circulating thyroid hormones and detailed assessment of reproductive organ parameters were added to the protocol. Reproductive hormones were suggested as additional parameters but they were deemed not informative in view of their large variability in untreated animals.

The developmental neurotoxicity guideline, accepted by OECD in 2007, has added the important aspect of behavioural effects of pre- and post-natal exposure to chemicals. This development arose from the notion that behavioural disorders in man such as anxiety, depression, phobias, autism and attention deficit hyperactivity disorder, which appear to show increasing prevalence in western societies, may have a perinatal origin ***(4,5)***. In the absence of causal inferences with respect to chemicals it seems nevertheless prudent to assess in a risk assessment whether such causal relations may exist.

A novel extended one-generation reproductive toxicity protocol was published in 2006 ***(6)***. Important aspects of the protocol are the addition of cohorts addressing developmental neurotoxicity and developmental immunotoxicity. The latter area also emanates from human data, in this case relating to allergies, eczema, asthma and autoimmune diseases, which show increasing prevalences and often have an early childhood onset ***(7-9)***. An elegant feature of the extended one-generation study protocol is that these additional parameters are studied without the need for additional animals to be entered into the study. Pups generated in the study that would not be studied in the classical two-generation study are now used for neurobehavioural or immune parameter assessment, conferring a substantial additional level of information to the test. Moreover, it is suggested that the second generation mating and offspring can be omitted from the test, as it will not give substantial additional information that will be crucial for hazard and risk assessment. This will reduce animal numbers in a single study from roughly 2600 to 1400, and could reduce overall animal use under REACH by an estimated 15%. The omission of the second generation mating and offspring has been disputed in the regulatory community. A retrospective analysis of close to 500 existing two-generation studies has shown that in none of these studies the data related to second generation mating and offspring affected risk assessment ***(10)***. Moreover, classification and labelling for reproductive toxicity was only very rarely influenced by the second generation mating and offspring in a two-generation study ***(11)***. From a regulatory perspective, the extended one-generation reproductive toxicity study has the advantages of additional parameters, increased numbers of animals being assessed for each parameter conferring higher statistical power and reduced animal use as compared to the two-generation study. OECD has adopted the draft guideline (OECD443) in July 2011 and governments are contemplating its inclusion in mandatory testing strategies, possibly as a replacement for the two-generation reproduction toxicity study (OECD416).

Reviewing the current range of OECD test protocols from the perspective of coverage of the reproductive cycle, with the generation studies covering (nearly) the entire cycle, each segment is also represented in a separate test with one exception. The early postnatal phase including puberty, briefly the juvenile period, is not represented in a separate test. Recent research has suggested that this developmental window, characterised by rapid growth and maturation of e.g., the brain, the immune system and the reproductive system, may be especially vulnerable to xenobiotic exposures ***(12,13)***. Although the generation studies include this period in their exposure scenario, these protocols do not allow pinpointing the window of vulnerability for the effects observed. Moreover, other effects may mask adverse effects specifically caused in the juvenile period. In the pharmaceutical area, juvenile toxicity study protocols have been subject to extensive study and debate over the years ***(14,15)***. This is especially important in view of drug safety for use in children. Similarly in the chemicals area, children have different exposure scenarios than adults, exemplified by the exposure to phthalates from sucking of baby toys containing high concentrations of plasticisers. Therefore, new initiatives are warranted towards a juvenile toxicity testing protocol that could be applied dependent on foreseen use of the chemicals of interest.

2. Alternative approaches

2.1 A brief state-of-the-art

The implementation of animal test protocols in the nineteen eighties has been accompanied by the development of a host of alternative methods to study adverse effects of chemicals on reproductive and developmental parameters. For example, rat whole embryo culture stems from the seventies ***(16)***, as does the rat limb bud organ culture ***(17)*** and rat limb bud and brain micromass was developed in the eighties ***(18)***. An elegant nonvertebrate alternative model used regeneration of polyps of *Hydra attenuata* from dissociated cells ***(19)***. Animal-free *in vitro* alternatives include those employing the proliferation of a human embryonic palatal mesenchymal cell line ***(20)***, the attachment of a mouse ovarian tumour cell line ***(21)*** and the differentiation of a neuroblastoma cell line ***(22)*** and a embryonal carcinoma cell line ***(23)***. Various overviews of methods have been published over the years ***(24)***. The predictability of alternative assays has been reviewed as well and has generally not passed the 80% mark ***(25)***.

The most extensive formal validation study in this area addressed whole embryo culture (WEC), micromass (MM) and the embryonic stem cell test (EST) ***(26)***. This validation study proved a great learning experience in terms of understanding the value of a study with a limited amount of diverse compounds in terms of extrapolation to the universe of chemicals. Subsequent application of the validated EST taught us that the 80% predictability was not reproduced with additional compounds ***(27)***. One of the issues underlying this discrepancy was in the mathematical prediction model used, which did not always appear to match the biology of the assay in terms of observed differentiation inhibition.

The three test systems have followed different routes after the completion of the combined validation study. MM has become less popular, perhaps because of its relatively low predictability in that study, but also due to its need for embryonic material, which does not make it a true animal-free alternative method. WEC also suffers from the need of embryo material and in addition, needs skilled personnel for isolation of viable early post-implantation embryos from their implantation sites in the uterus. Nevertheless, WEC has the advantage of employing the development of intact embryos as well as covering a critical period in organogenesis, which has been instrumental in addressing dysmorphogenetic effects of a diversity of compounds. Recent comparative studies of gene expression changes in WEC versus *in vivo* have confirmed large similarities between *in vitro* and *in vivo* development at the molecular level, confirming the relevance of the assay ***(28)***.

The EST has continued to be a popular test subject to further optimisation, especially through the addition of alternative pathways of differentiation and the addition of molecular markers of effects. Besides cardiac muscle differentiation, other cell types such as neurons ***(29-31)***, osteoblasts ***(32)***, adipocytes ***(33)*** and hepatocytes ***(34)*** have been generated in dedicated differentiation protocols, providing additional opportunities for testing the interference of differentiation pathways with chemical exposures. The addition of transcriptomics approaches to measure differential gene expression, both as influenced by the differentiation process as well as due to chemical exposure, has proven informative and may enhance the predictability of EST assays ***(35)***. The extension of this concept to human embryonic stem cell lines is still in its infancy, but has opened a new realm of options that bypasses the need for interspecies extrapolation ***(36)***.

Additional assays have surfaced at both ends of the scale of complexity. Single end-point assays such as those for a host of receptor binding and activation assays and enzyme activity modulation assays have been developed

and applied to alternative developmental toxicity testing. These assays can be carried out in a high-throughput fashion, allowing vast numbers of compounds to be tested in a large battery of assays. The USEPA ToxCast programme has the most widely developed alternative test database, with around 300 assays and one thousand compounds tested in their first phase, but this is only the beginning of a large-scale project ***(37,38)***. The ambitious aim is to derive a developmental toxicity signature that can be used for prediction on the basis of this *in vitro* battery only. The challenge clearly is in bridging the gap between the large group of individual single end-point assays and the integrated complexity of embryo-fetogenesis in time and space.

At the other end of the complexity scale, the zebra fish embryotoxicity assay gains popularity ***(39,40)***. It offers the study of a complete vertebrate embryo in its natural environment (the egg), which develops within one week from fertilisation to a hatched, free-swimming fish. During the first 96 hours of development the embryo is dependent on its yolk for feeding and on that basis is not legally considered an experimental animal. Therefore, this test is formally animal free. The advantages of the system include easy availability and handling, the possibility of observing apical effects on the intact embryo, easy observation of the transparent embryo during development and simple exposure to compounds via the surrounding water. In addition, the genetic code of the zebra fish has been fully mapped and a multitude of mutants, siRNAs and microRNAs is available, allowing dedicated molecular studies. Disadvantages of the system are the low penetration of the chorion by certain classes of compounds, which can only partly be overcome by dechorionation or microinjection. Furthermore, the throughput of the system is evidently lower than that of molecular assays. Clearly, the combination of single end-point assays with more integrative assays in testing strategies could combine the advantages of both types of approach.

2.2 Toxicity testing in the 21st century

The call for alternative approaches towards hazard assessment of xenobiotics has been stimulated by ethical considerations of animal use as well as by the wish for a more mechanistic basis for human hazard assessment. The extrapolation of findings in animal studies to actual hazard and risk in man requires knowledge about the relevance of toxicological mechanisms of action in both species. Ideally, toxic effects in man should inform about human hazard and risk but experimental studies in man are ethically unacceptable. Two possible ways to bridge the mechanistic interspecies gap can be envisaged. First, mechanisms of action can be elucidated in dedicated *in vitro* assays, preferably using human material and secondly, hazard identification could start from the human disease perspective, relating diseases to actual human exposures. Both approaches have been contemplated and promoted in the US National Academy of Sciences report on “Toxicity Testing in the 21st Century, a Vision and a Strategy” ***(41)***. The success of such approaches critically depends on our ability to identify the essential mechanistic elements of toxicity pathways and to rebuild them in *in vitro* models and on the possibility to conclude about cause and effect relationships between human exposure and disease. Both of these aspects provide significant challenges for research.

2.3 Endocrine disruption

The issue of endocrine disruption has greatly stimulated research into alternative methods for assessing endocrine activities of xenobiotics. The most widely accepted definition of an endocrine disrupter includes those compounds that can cause an adverse health effect in an intact animal through modulation of an endocrine mechanism ***(42)***. This requires that an adverse health effect is shown in an intact animal, which in principle cannot be detected by an *in vitro* alternative test. Yet even in *in vivo* tests, the chicken and egg question about the possible causal relation between endocrine modulation and adverse health effects is not always easily answered. Many chemicals were shown to have at least some activity in (reproductive) hormone receptor binding and activation assays. Although their potency often appears to be very low, such positive tests have readily resulted in earmarking such chemicals as potential endocrine disrupters. Consideration of *in vivo* exposure scenarios and kinetics of such chemicals in the body often indicated that actual risks of even high accidental exposures would probably be negligible. Such conclusions were drawn in view of physiologic background levels of natural hormones and homeostatic properties of the various hormonal feedback loops in the intact individual, such as the hypothalamic-pituitary-adrenal axis ***(e.g., 43).***

In vivo single endocrine end-point assays have regained attention as well, such as the uterotrophic assay for estrogenicity ***(44)*** and the Hershberger assay for androgenicity ***(45)***. These *in vivo* assays have recently been upgraded to OECD guideline assays, secondary to endocrine disruptor testing programmes. However, their actual use in testing strategies is subject to the discussion about their added value over *in vitro* receptor activation assays. As to compound kinetics in the animal, these assays offer only limited additional value if exposure is via the subcutaneous, intravenous or intraperitoneal routes, which are less relevant for actual human exposure. In addition, these assays are often performed in ovariectomised and castrated animals. This design makes it difficult to extrapolate findings to the intact animal, which has the complete homeostatic feedback loops in place that offer protection especially for exogenous endocrine exposures. The extra animal use over *in vitro* assays further complicates the discussion about the added value of these assays.

The endocrine disruptor issue has taught us a number of lessons, a number that may well increase in the near future. Let us here restrict ourselves to remarking that the interpretation and extrapolation of *in vitro* assay results should be done carefully, considering the potency of the compound tested, the abundant kinetic differences between *in vitro* and *in vivo* exposures, and considering the relevance of the *in vitro* end-point measured for the *in vivo* situation in terms of adverse health effects.

2.4 Issues with alternative *in vitro* assays

Alternative assays to animal testing are of interest for two main reasons. First of all, animal testing is expensive and time-consuming and it has ethical issues. Second, the relevance to man of animal testing is generally not well understood. Alternative methods offer the possibility to use human biological material, such as established cell lines from human tissues. The possibility for more detailed mechanistic analysis of compound effects in *in vitro* assays can help to improve extrapolation of effects between species. In addition, they may enable a more scientifically based assessment of hazard and risk for the human population. The intrinsic properties of *in vitro* alternative assays offer advantages and disadvantages, which need careful consideration to determine their optimal application in hazard and risk assessment.

First and foremost, alternative assays are largely reductionistic, especially in a complex area such as reproductive toxicology. For example, the embryonic stem cell test offers insight into compound effects on cardiac cell differentiation and possibly, on other differentiation pathways that influence the primary end-point observed. Omics approaches may enhance the readout of the system, allowing a more detailed assessment of compound effects. Nevertheless, important aspects of development such as three-dimensional pattern and organ formation, and the attainment of functionality beyond cardiac muscle cell contraction cannot be studied in this assay. This has important consequences both for the interpretation of assay results as well as for its use in a testing strategy.

As effect parameters studied in *in vitro* assays become more refined, the question becomes more important of how to interpret the findings in terms of toxicity. For example, in transcriptomic experiments, very low exposures in the EST which do not affect cellular end points of proliferation and differentiation can be shown to affect gene expression ***(46)***. The question arises at what level of response the observed effect reaches the level of adversity. Evidently, many physiologic responses are benificial, neutralising the hazard by homeostatic control and therefore not all detectable responses should be characterised as toxic or adverse. It may not even be possible to answer this question on the level of an individual test, but this may require a more integrated approach using weight of evidence over a combination of results from different assays.

As to the interpretation of individual assay results, a reductionistic assay will by definition not predict 100% of developmental toxicants tested, as for at least some of them the mechanism of developmental toxicity will not be covered within the assay. Therefore, a validation exercise with a variety of compounds with unknown mechanisms of developmental toxicity has only limited value if only used to derive an overall predictability rate of a single assay. It is more useful to elucidate the applicability domain of the assay in terms of the mechanisms of development covered, and to validate that aspect by testing compounds that do or do not affect the applicability domain. For single end-point assays such as specific receptor activation assays, this exercise is relatively straightforward. For more complex assays such as those involving embryonic cell differentiation, the understanding of the applicability domain is more complex, as extensive research with the embryonic stem cell test has taught us ***(27,47)***. Whole embryo cultures are probably more straightforward in terms of applicability domain, as they involve the entire embryo in a limited window of development, but such assays are complex and not animal-free.

Applicability domain can also be defined by the chemical classes for which an assay is predictive of their developmental toxicity ***(48)***. Kistler ***(49)*** showed that the MM nicely predicted the relative teratogenicity potential of a series of retinoids. Such an assay can be instrumental in designing new representatives of chemical classes with more favourable characteristics in terms of pharmacological *versus* toxic characteristics. Concurrently, it should be kept in mind that additional toxic effects may arise in new compounds, requiring more elaborate testing in a variety of test systems. This leads to the use of tiered or battery approaches, combining a series of complementary assays covering all important mechanisms of development. The combined applicability domain of a group of diverse alternative assays may improve overall predictability for a large variety of chemicals. It will be important to develop such batteries of tests in order to enhance the overall predictability of alternative approaches.

Finally, an often neglected issue with alternative *in vitro* assays is the aspect of compound kinetics. *In vitro* exposures are obviously very different from *in vivo* exposures. Absorption, distribution, metabolism and excretion are

largely absent in the *in vitro* tests. Extrapolation from *in vitro* effective concentrations to *in vivo* effective dosages and their changes with time after exposure requires internal target organ exposure modelling ***(50)***. Moreover, exposure level over time may be largely stable *in vitro* whereas *in vivo* peak exposures may occur, complicating comparison. Developmental toxicity is a dose-dependent phenomenon, therefore no compound can simply be scored as a positive or a negative, but potency is critically important. The recent ILSI-HESI-DART approach to define “exposures” rather than compounds as developmental toxicants is taking account of this notion ***(51)***. An “exposure” is defined as a compound in combination with its internal exposure level. These “exposures” can be used to validate an *in vitro* assay, taking account of compound potency in the determination of predictability of an assay.

3. Strategy for innovation

3.1 Introduction

The above considerations provide ample opportunities for innovations in testing strategies in reproductive toxicology. Clearly, the experience with standardised regulatory animal testing over the last three decades has provided us with a wealth of relevant information, which could be mined to elucidate patterns of toxicity that can inform about ways towards innovation. Similarly, human exposure and disease data can be employed for that aim. When critical end-points and modes of action of toxicity are mapped, underlying mechanisms may be elucidated and dedicated alternative *in vitro* tests can then be designed. Although the latter is easier said than done, it is essential for the regulatory acceptability of alternative assays. As past experience and common sense has taught us, individual *in vitro* assays cannot be expected to fully predict generalised complex apical end-points such as developmental toxicity. Therefore, testing strategies combining all essential mechanistic components of hazard identification should be designed. However, it would be unrealistic to presume that alternative testing strategies will fully replace animal testing in the near future. Human hazard identification is too critically important to allow a rapid transition to as yet uncertain alternative approaches, even when considering the fact that the animal as gold standard for human hazard assessment is far from optimal. However, alternative testing strategies can be instrumental in reducing and refining animal testing to limited use as a last resort only, also significantly reducing testing time and cost. At the same time, alternative approaches will enhance human hazard identification, basing it on mechanistic *in vitro* data that can be extrapolated to human hazard. Several current developments in this area show promise, such as those reviewed below.

3.2 Database analysis

Database analyses are gaining interest in view of their informative properties. Large integrated databases such as the ToxRefDB at USEPA allow complex analyses comparing, e.g., different study protocol outcomes as well as relative parameter sensitivity ***(37)***. In reproductive and developmental toxicology, the added value of the generation study in the presence of subchronic toxicity test data ***(52)*** and the relative contribution of rabbit *versus* rat prenatal developmental toxicity studies ***(53)*** to hazard assessment have been studied using existing data from guideline-based studies. In addition, the impact of the second generation mating and offspring parameters in the two-generation reproductive toxicity study has been analysed on the basis of an interagency

database of close to 500 existing studies ***(10)***. Such analyses can be instrumental in reducing animal use to the necessary minimum. Moreover, they help identify those end-points within these complex studies that are the most informative in view of subsequent hazard identification. These selected end-points merit representation in dedicated *in vitro* assays whenever possible. One area of specific attention is the testis, which often showed specific sensitivity in animal studies but for which currently no satisfactory alternative assay exists. More general end points that are often affected in reproductive toxicity studies, such as pup weight and litter size, are conceptually difficult to mimic *in vitro*, and in those cases extrapolation of mechanistic information from *in vitro* studies may eventually be of help. Apart from animal study databases, the use of human data may improve human hazard identification. The US-NAS paper has pointed towards the identification of human exposure and disease relationships on the basis of patient data ***(41)***. Although causality is often difficult to prove, human data are important in identifying toxicity end-points that merit detailed assessment in hazard identification. Increased prevalences of immune-related diseases and neurodevelopment-related conditions in western populations have been among the reasons for suggesting specific end-points for developmental immunotoxicity and developmental neurotoxicity in the extended one-generation reproductive toxicity study ***(6)***. Thus, both animal and human databases can assist in enhancing adverse health effects detection, which could help improve efficiency in regulatory toxicology, reducing animal use and indicating end-points for the development of alternative assays.

3.3 Alternative methods

The development and refinement of alternative methods has gained tremendously from rapid innovations in molecular biology and cell culture techniques. It is now possible to refine traditional end-points in *in vitro* models, such as cell proliferation, cell death and cell differentiation inhibition by studying molecular end-points through, e.g., transcriptomics, proteomics and metabolomics. Such end-points have a very high informative value which, once well understood, may increase the predictability of *in vitro* assays. For instance, in EST the traditional end-point of cardiac differentiation, observed by counting contracting cardiac muscle cell foci under the microscope, can now be enhanced by transcriptomics analysis of the entire genome, allowing a depth of analysis way beyond cardiac myocyte differentiation ***(54)***. In addition, in whole embryo culture, early transcriptomics analysis may enable early prediction of morphologic effects on, e.g., palatal closure occurring six days beyond the culture period and therefore undetectable by morphological observation during culture ***(28)***. Three dimensional culture techniques are becoming available which enable the reconstruction *in vitro* of functional organs such as the thyroxin producing thyroid ***(55)***. Eventually, three dimensional tissue engineering approaches may also enable us to rebuild the functional sperm-producing testis to allow *in vitro* testing of compounds affecting spermatogenesis at the level of the testis. Finally, culture techniques enable the transition from rodent to human cells as the basis for *in vitro* assays, such as currently ongoing with human embryonic stem cells. The latter would allow direct testing in assays developed from biological material of the species relevant for hazard and risk assessment. In spite of these accumulating innovations, the implementation of alternatives has lagged behind. This is partly due to the uncertainty about applicability domains as discussed earlier, which makes it difficult to determine the optimal use of these assays. Their eventual employment will probably be in a combination of assays in a well-balanced testing strategy, as discussed below.

3.4 Testing strategies

Given the complexity of reproductive toxicity, it is logical to propose that at least combinations of complementary alternative assays will be needed to cover the entire spectrum of mechanisms of toxicity *in vitro*. In a tiered and/or battery approach, assays could produce a multilayered filter for detecting toxic properties, pragmatically starting with simple high throughput assays and ending in the more apical and complex low throughput assays. This *in vitro* filter should be integrated with all additional available knowledge on the chemicals under study. The OECD framework for reproductive toxicity testing proposes several levels of assessment, starting with non-testing information such as chemical structure, physical properties and category approaches or grouping to anticipate possible toxic properties of chemicals ***(56)***. Following this non-testing assessment, various levels of *in vitro* assays can be envisaged, which may be tailored on the basis of pre-existing information on the chemical. Next, the first *in vivo* step could be single end-point assays such as uterotrophic and Hershberger assays, although their added value can be disputed (as discussed above). The classical apical generation studies and developmental toxicity studies in animals still would serve as the final filters for cases where earlier steps have not given the required definitive answers about hazard.

Clearly the challenge is in designing the multi-assay *in vitro* filter to optimally detect toxic moieties. The largest study to date aimed at defining an *in vitro* test battery is being performed by the USEPA ToxCast project, which is testing thousands of compounds in hundreds of high throughput single end-point *in vitro* assays ***(57)***. The data analysis in this project aims at defining a minimal essential set of assays that adequately predict developmental toxicity. So far, the predictability achieved in these analyses has been promising but has not reached the 80% level that has been so familiar in validation studies of a number of individual alternative assays in the past ***(25)***. A pragmatic proof-of-principle approach was taken by the Reprotect project ***(58)***, in which ten available standardised assays were used to predict the developmental toxicity of ten blinded compounds. This study was promising in that it gave favourable results for the compounds tested, although it should be noted that all but one of the selected compounds had distinct modes of action detectable in one of the assays employed. The single compound not detected in that study was a testicular toxicant, an end-point for which the small test battery employed did not have a representative test available. A similar study is currently carried out within the ChemScreen project, which aims at generating a pragmatic testing strategy on the basis of current knowledge ***(59)***. The latter exercise will show us where we are in terms of achievements as well as indicate knowledge gaps as to mechanisms of toxicity and related tests required to fill all the current gaps in the alternative assay filter to be designed.

Computational toxicology is being explored as the tool that could integrate all information from the molecular to the organism level. The virtual embryo and virtual liver projects at USEPA are examples which could lead to full *in silico* predictive models for toxicity prediction ***(60)***. These projects aim to functionally integrate the single end-point *in vitro* assays together on the mechanistic level, which may improve the interpretation of the assays in terms of hazard identification and enhancing predictability on a higher level of integration. This approach should additionally allow integrating all areas of toxicity testing. For example, potent genotoxic compounds could be filtered out at an early stage and need not enter reproductive toxicity testing. Such integration beyond the current subdisciplines of toxicology could probably provide the largest gain in efficiency in hazard identification. These approaches, tackling innovation of hazard identification from different angles, should

ultimately converge at some point in time to combine mechanisms of toxicity, representative *in vitro* assays, and a logical testing battery and/or tier to comprehensively detect compounds hazardous for man and their potency in the absence of animal testing.

References

1. EU REACH 2011. REACH legislation for Registration, Evaluation, Authorisation and Restriction of Chemicals: http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm, accessed 24 August 2011.

2. OECD 2011. Test guidelines program: http://www.oecd.org/department/0,2688,en_2649_34377_1_1_1_1_1,00.html, accessed 24 August 2011.

3. EU GHS 2011. Globally Harmonised System for classification, labelling and packaging of substances and mixtures: http://ec.europa.eu/enterprise/sectors/chemicals/classification/index_en.htm, accessed 24 August 2011.

4. Jurewicz J, Hanke W. (2008) Prenatal and childhood exposure to pesticides and neurobehavioral development: review of epidemiological studies. Int J Occup Med Environ Health.21(2):121-32.

5. Mill J, Petronis A. (2008) Pre- and peri-natal environmental risks for attention-deficit hyperactivity disorder (ADHD): the potential role of epigenetic processes in mediating susceptibility. J Child Psychol Psychiatry 49(10):1020-30.

6. Cooper RL, Lamb JC, Barlow SM, et al., (2006) A tiered approach to life stages testing for agricultural chemical safety assessment. Crit Rev Toxicol.36(1):69-98.

7. Timmer A. (2003) Environmental influences on inflammatory bowel disease manifestations. Lessons from epidemiology. Dig Dis.21(2):91-104.

8. Woodruff TJ, Axelrad DA, Kyle AD, et al. (2004) Trends in environmentally related childhood illnesses. Pediatrics 113(4 Suppl):1133-40.

9. Pearce N, Douwes J. (2006) The global epidemiology of asthma in children. Int J Tuberc Lung Dis.10(2):125-32.

10. Piersma AH, Rorije E, Beekhuijzen ME, et al., (2011) Combined retrospective analysis of 498 rat multi-generation reproductive toxicity studies: On the impact of parameters related to F1 mating and F2 offspring. Reprod Toxicol. 31(4):392-401.

11. Rorije E., A. Muller, M. E.W. Beekhuijzen, et al., (2011) On the impact of second generation mating and offspring in multi-generation reproductive toxicity studies on Classification and Labelling of Substances in Europe. Regul. Pharmacol. Toxicol., in press.

12. Tonk EC, de Groot DM, Penninks AH, et al. (2010) Developmental immunotoxicity of methylmercury: the relative sensitivity of developmental and immune parameters. Toxicol Sci. 117(2):325-35.

13. Tonk EC, de Groot DM, Penninks AH, et al., Developmental immunotoxicity of di-n-octyltin dichloride (DOTC) in an extended one-generation reproductive toxicity study. Toxicol Lett. 204(2-3):156-63.

14. Cappon GD, Bailey GP, Buschmann J, et al., (2009) Juvenile animal toxicity study designs to support pediatric drug development. Birth Defects Res B Dev Reprod Toxicol.86(6):463-9.

15. Shimomura K. (2011) The value of juvenile animal studies: a Japanese industry perspective. Birth Defects Res B Dev Reprod Toxicol. 2011 May 18. [Epub ahead of print]

16. New DAT. (1978) Whole-embryo culture and the study of mammalian embryos during organogenesis. Biol Rev Camb Philos Soc. 53(1):81-122.

17. Agnish ND, Kochhar DM. (1976) Direct exposure of mouse embryonic limb-buds to 5-bromodeoxyuridine in vitro and its effect on chondrogenesis: increasing resistance to the analogue at successive stages of development. J Embryol Exp Morphol. 36(3):639-52.

18. Flint OP, Orton TC, Ferguson RA. (1984) Differentiation of rat embryo cells in culture: response following acute maternal exposure to teratogens and non-teratogens. J Appl Toxicol. 4(2):109-16.

19. Johnson EM. (1980) A subvertebrate system for rapid determination of potential teratogenic hazards. J Environ Pathol Toxicol. 4(5-6):153-6.

20. Welsch F, Stedman DB, Willis WD, et al. (1986) Karyotype, growth, and cell cycle analysis of human embryonic palatal mesenchymal cells: relevance to the use of these cells in an in vitro teratogenicity screening assay. Teratog Carcinog Mutagen. 6(5):383-92.

21. Steele VE, Morrissey RE, Elmore EL, et al. (1988) Evaluation of two in vitro assays to screen for potential developmental toxicants. Fundam Appl Toxicol. 11(4):673-84.

22. Mummery CL, van den Brink CE, van der Saag PT, et al. (1984) A short-term screening test for teratogens using differentiating neuroblastoma cells in vitro. Teratology. 29(2):271-9.

23. Piersma AH, Haakmat AS, Hagenaars AM. (1993) In vitro assays for the developmental toxicity of xenobiotic compounds using differentiating embryonal carcinoma cells in culture. Toxicology in vitro, 7: 615-621.

24. Piersma AH. (2006) Alternative methods for developmental toxicity testing. Basic Clin Pharmacol Toxicol. 98(5):427-31.

25. Brown NA. (1987) Teratogenicity testing in vitro: status of validation studies. Arch Toxicol Suppl. 11:105-14.

26. Genschow E, Spielmann H, Scholz G, et al. (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):151-76.

27. Marx-Stoelting P, Adriaens E, Ahr HJ, et al. (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):313-28.

28. Robinson JF, Theunissen PT, van Dartel DA, et al. (2011) Comparison of MeHg-induced toxicogenomic responses across in vivo and in vitro models used in developmental toxicology. Reprod Toxicol.32(2):180-8.

29. Bain G, Ray WJ, Yao M, et al. (1996) Retinoic acid promotes neural and represses mesodermal gene expression in mouse embryonic stem cells in culture. Biochem Biophys Res Commun 223(3):691–4.

30. Okabe S, Forsberg-Nilsson K, Spiro AC, et al. (1996) Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. Mech Dev 59(1):89–102.

31. Theunissen PT, Schulpen SH, van Dartel DA, et al. (2010) An abbreviated protocol for multilineage neural differentiation of murine embryonic stem cells and its perturbation by methyl mercury. Reprod Toxicol. 29(4):383-92.

32. zur Nieden NI, Kempka G, Ahr HJ. (2003) In vitro differentiation of embryonic stem cells into mineralised osteoblasts. Differentiation 71(1):18-27.

33. Dani C, Smith AG, Dessolin S, et al. (1997) Differentiation of embryonic stem cells into adipocytes in vitro. J Cell Sci. 110:1279-85.

34. Hamazaki T, Iiboshi Y, Oka M, et al. (2001) Hepatic maturation in differentiating embryonic stem cells in vitro. FEBS Lett. 18;497(1):15-9.

35. van Dartel DA, Piersma AH (2011). The embryonic stem cell test combined with toxicogenomics as an alternative testing model for the assessment of developmental toxicity. Reprod Toxicol.32(2):235-44.

36. Stummann TC, Hareng L, Bremer S. (2009) Hazard assessment of methylmercury toxicity to neuronal induction in embryogenesis using human embryonic stem cells. Toxicology. 29;257(3):117-26.

37. Knudsen TB, Martin MT, Kavlock RJ, et al. (2009) Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. Reprod Toxicol. 28(2):209-19.

38. Reif DM, Martin MT, Tan SW, et al. (2010) Endocrine profiling and prioritisation of environmental chemicals using ToxCast data. Environ Health Perspect. 118(12):1714-20.

39. Scholz S, Fischer S, Gündel U, et al. (2008) The zebrafish embryo model in environmental risk assessment--applications beyond acute toxicity testing. Environ Sci Pollut Res Int. 15(5):394-404.

40. Hermsen SA, van den Brandhof EJ, van der Ven LT, et al. (2011) Relative embryotoxicity of two classes of chemicals in a modified zebrafish embryotoxicity test and comparison with their in vivo potencies. Toxicol In Vitro.25(3):745-53.

41. US-NAS (2007) Toxicity Testing in the 21st Century: A Vision and a Strategy. National Academies Press, Washington DC.

42. EC (1996) Report of Proceedings of European Workshop on the Impact of Endocrine Disruptors on Human Health and Wildlife, Weybridge, UK. http://ec.europa.eu/environment/endocrine/documents/reports_en.htm, accessed 24 August 2011.

43. Hengstler JG, Foth H, Gebel T, et al. (2011) Critical evaluation of key evidence on the human health hazards of exposure to bisphenol A. Crit Rev Toxicol. 41(4):263-91.

44. Tyl RW, Marr MC, Brown SS, et al. (2010) Validation of the intact rat weanling uterotrophic assay with notes on the formulation and analysis of the positive control chemical in vehicle. J Appl Toxicol. 30(7):694-8.

45. Freyberger A, Schladt L. (2009) Evaluation of the rodent Hershberger bioassay on intact juvenile males--testing of coded chemicals and supplementary biochemical investigations. Toxicology. 262(2):114-20.

46. van Dartel DA, Pennings JL, de la Fonteyne LJ, et al. (2011) Concentration-dependent gene expression responses to flusilazole in embryonic stem cell differentiation cultures. Toxicol Appl Pharmacol. 251(2):110-8.

47. Paquette JA, Kumpf SW, Streck RD, et al. (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):104-11.

48. Hartung T, Bremer S, Casati S, et al. (2004) A modular approach to the ECVAM principles on test validity. Altern Lab Anim.32(5):467-72.

49. Kistler A. (1987) Limb bud cell cultures for estimating the teratogenic potential of compounds. Validation of the test system with retinoids. Arch Toxicol. 60(6):403-14.

50. Louisse J, de Jong E, van de Sandt JJ, et al. (2010) The use of in vitro toxicity data and physiologically based kinetic modelling to predict dose-response curves for in vivo developmental toxicity of glycol ethers in rat and man. Toxicol Sci.118(2):470-84.

51. Daston GP, Chapin RE, Scialli AR, et al. (2010) A different approach to validating screening assays for developmental toxicity. Birth Defects Res B. 89(6):526-30.

52. Janer G, Hakkert BC, Piersma AH, et al. (2007) A retrospective analysis of the added value of the rat two-generation reproductive toxicity study versus the rat subchronic toxicity study. Reprod Toxicol.24(1):103-13.

53. Janer G, Slob W, Hakkert BC, et al. (2008) A retrospective analysis of developmental toxicity studies in rat and rabbit: what is the added value of the rabbit as an additional test species? Regul Toxicol Pharmacol. 50(2):206-17.

54. Pennings JL, van Dartel DA, Robinson JF, et al. (2011) Gene set assembly for quantitative prediction of developmental toxicity in the embryonic stem cell test. Toxicology. 284(1-3):63-71.

55. Toda S, Aoki S, Suzuki K, et al. (2003) Thyrocytes, but not C cells, actively undergo growth and folliculogenesis at the periphery of thyroid tissue fragments in three-dimensional collagen gel culture. Cell Tissue Res.312(3):281-9.

56. OECD 2002. Conceptual framework for testing and assessment of potential endocrine disruptors.
http://www.oecd.org/document/58/0,3343,en_2649_34377_2348794_1_1_1_1,00.html, accessed 24 August 20911.

57. Martin MT, Knudsen TB, Reif DM, et al. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. Biol Reprod. 85(2):327-39.

58. Schenk B, Weimer M, Bremer S, et al. (2010) The ReProTect Feasibility Study, a novel comprehensive in vitro approach to detect reproductive toxicants. Reprod Toxicol. 30(1):200-18.

59. van der Burg B, Kroese ED, Piersma AH. (2011) Towards a pragmatic alternative testing strategy for the detection of reproductive toxicants. Reprod Toxicol. 31(4):558–61.

60. Kavlock R, Dix 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.

National Insitute for Public Health
and the Environment
P.O. Box 1 | 3720 BA Bilthoven
www.rivm.com