A Critique of the EC's Expert (Draft) Reports on the Status of Alternatives for Cosmetics Testing to Meet the 2013 Deadline

Katy Taylor, Carlotta Casalegno and Wolfgang Stengel

European Coalition to End Animal Experiments (ECEAE), London, UK

Summary

The 7th Amendment to the EU's Cosmetic Directive (now recast as Regulation 1223/2009) bans the testing of cosmetic ingredients and products on animals, effective 2009. An extension until 2013 was granted, for marketing purposes only, for three endpoints: repeated dose, toxicokinetics, and reproductive toxicity. If the European Commission determines that alternatives for these endpoints are not likely to be available, it can propose a further extension. To this end, the Commission has instructed experts to produce reports on the status of alternatives for the 2013 deadline.

We criticized the draft reports on a number of issues. First, the experts fell into the "high fidelity fallacy trap," i.e. asserting that full replication of the in vivo response, as opposed to high predictivity, is required before an animal test can be considered useful for regulatory purposes. Second, the experts' reports were incomplete, omitting various methods and failing to provide data on the validity, reliability, and applicability of all the methods discussed, regardless of whether the methods were in vivo, in vitro, or in silico.

In this paper we provide a summary of our criticisms and provide some of the missing data in an alternative proposal for replacement of animal tests by 2013. It is our belief that use of the Threshold of Toxicological Concern (TTC) will be a useful method to mitigate much animal testing. Alternative approaches for carcinogenicity and skin sensitization could be considered sufficient in the very near future, even though these tests are not listed under the 2013 extension. For repeated dose, toxicokinetics, and reproductive toxicity a combination of in vitro methods may be able to provide appropriate protection for consumers, especially when viewed in the context of the poor predictivity of the animal models they replace. We hope the revised report will incorporate these comments, since a more thorough and positive review is required if the elimination of animal testing for cosmetics in Europe and beyond is to be achieved.

Keywords: Cosmetics, expert report, alternatives, in vitro, TTC

1 Introduction

The Cosmetics Directive (Directive 76/768/EEC), now recast as Regulation 1223/2009, bans testing on animals for cosmetic purposes as of March 2009. In addition the "marketing" (i.e. import and sale) of products and ingredients tested on animals outside Europe also is prohibited after that date. A postponement of the marketing ban until 2013 was provided for three endpoints however – toxicokinetics, repeated dose, and reproductive toxicity. At the time, these tests were considered harder to replace.

Under the terms of the Cosmetics Directive, the European Commission shall publish a proposal for an extension of this deadline if it decides that alternatives for these three tests will not be available by this date. As part of this process, the Commission gathered experts from across Europe in May 2010 to produce a report on the status of alternatives for cosmetics testing. The aim of the experts' report was to evaluate whether alternatives would be available by 2013 and, if not, to establish recommendations and a timeline for complete replacement. In July 2010, a draft version of the report was made available for public comment, comprising five chapters covering five endpoints: repeated dose, skin sensitization, carcinogenicity, toxicokinetics, and reproductive toxicity (http://ec.europa.eu/consumers/sectors/cosmetics/documents/public_consultation/index_en.htm). The reports covered how information on the endpoint is currently derived, (i.e. Scientific Committee on Consumer Safety (SCCS) requirements for *in vivo* tests) and then summarized the various alternative approaches, including *in vitro* methods, QSARs (Quantitative Structure Activity Relationships), and

Received February 3, 2011; accepted in revised form March 28, 2011.

other *in silico* methods. A table was consistently used to list the methods by mechanism of action, area of application, and "status" (i.e., in research and development, optimization, prevalidation, validation or regulatory acceptance).

We submitted extensive comments on the draft chapters as the ECEAE (the European Coalition to End Animal Experiments). The ECEAE was one of the leading pan-European organizations that campaigned for an end to cosmetics testing on animals in Europe under the Cosmetics Directive. We therefore have an interest in ensuring the bans are upheld, for ethical, if not for scientific reasons. Regardless of the ethical dimension to this debate, it was our opinion that the experts' reports were significantly scientifically flawed in a number of ways. This is a concern to us since we wish the debate to be founded on the best possible scientific evidence and assessment. We have had a recent poor experience with a similar expert report that is currently being investigated by the EU Ombudsman for lack of balance (Bailey and Taylor, 2009), and we would like that to be avoided in this case.

First, we contend that the experts applied the wrong legal test, asking whether it was possible to replicate the animal model in full, and not whether alternatives could *predict* human effects reliably. Given this "ultimate challenge" approach, it is not surprising that the experts believed the 2013 deadline would not be met for any of the five tests and could not, in most cases, offer a rational timescale for when they would be.

Irrespective of differences of opinion regarding the correct approach, we felt the reports, on the whole, lacked a proper evaluation of the status of the alternatives. Without this, we feel it is impossible to assess where the weaknesses are and what additional research is needed. A proper evaluation, in our opinion, includes assessment of the reliability of the method, its accuracy (including concordance with *in vivo* or gold standard methods, sensitivity and specificity), applicability domain (based on the known mechanism of action, range of substances used in any evaluation or known physical or biological limitations of the test) and, finally, the availability of the method. The point at which methods would, in all likelihood, be considered adequate by the regulators (i.e. the SCCS), if they applied the correct legal test, should have been given and all methods (whether *in vivo* or *in vitro*) rated against this.

The reports also were inconsistent in their approach, with some chapters (e.g., the one on toxicokinetics) adopting a more proactive, forward-thinking strategy. Some included concepts such as the Threshold of Toxicological Concern (TTC), an approach that can mitigate some tests, while others did not.

Finally, the inclusion of two additional endpoints (skin sensitization and carcinogenicity) in the reports was done under instruction from the Commission, with the implication that they too fall under the 2013 deadline, which is not consistent with the Cosmetics Directive and is of grave concern to us.

What follows is a summary of our comments on the draft reports, providing examples of where each chapter could have been improved. We looked at a number of key measures of quality in the chapters, including neutrality, completeness, use of quantitative measures, relevance of information to the cosmetics sector, as well as consistency across chapters for these measures. We include comments from experts that assisted with our submission to the Commission. In addition, for the purposes of this paper, for each endpoint, we offer a stepwise approach to testing, incorporating the TTC approach and validated methods that could be considered as adequate. We do this to stimulate debate on the adequacy of alternative methods and to provide an example of a basis from which we would have liked the experts to start. At the time of writing, the experts' final report has yet to be published, so our comments remain relevant to the draft version only. It is hoped that some of the comments included here have already been taken on board in their revisions and will be seen in the final report. Since there has been considerable delay from the publication of the draft report (July 2010) to the appearance of the final report (still not available in March 2011), we think there is value in offering our position on the status of alternative methods in advance of its publication.

2 General comments on experts' approach

The Commission, in instructing the experts, obviously gave them terms of reference (Anon, 2010). The principle aim of their reports was to evaluate the "state of play" of alternatives for the 2013 deadline, and "that the state of play must be as neutral as possible." From the outset, however, the assumption was that alternatives would not be available: "It should provide a wide and objective overview on the technical difficulties in complying with the ban in relation to tests evaluating repeated-dose toxicity, reproductive toxicity and toxicokinetics, in particular those for which there are no replacement methods or strategies yet under consideration. It should summarize the status and prospects of alternative methods and a scientifically sound estimate of the time necessary to achieve full replacement of animal testing for the above mentioned complex endpoints."

What was not made clear, however, was that the evaluation of what could be considered a "replacement" should refer to methods that can be used for the purposes of the Cosmetics Directive, i.e., for regulatory purposes. It appears that the approach taken throughout the reports was broader than this, i.e., to examine what is required before the entire mechanism of action of a particular toxic reaction can be understood and modeled. This falls into the high-fidelity fallacy trap recognized in 1959 by Russell and Burch (1959), and also known as the "uncertainty paradox" (Schaafsma et al., 2009). This is the assumption that *in vivo* animal models are automatically superior models of the human response and that in order to be useful, non-animal methods must replicate the *in vivo* animal model in full.

This approach fails to recognize that not all aspects of the mechanism of action need to be covered by a model for it to be highly predictive (and therefore useful for regulatory purposes). If the requirement of any given replacement method is to fully replicate the *in vivo* response, then this obviously will result in inordinately long timescales for their development. It is also an unfair comparison when animal models, while providing a picture of the "whole body response" (which *in vitro* methods cannot), nonetheless are *the wrong* body and therefore have unavoidable errors due to species differences. As a result, ani-

mal models themselves are not accurate predictors of the human response. For example, extrapolation factors have to be added to the results of animal tests to account for inter-species differences. And yet the experts did not evaluate the evidence for the validity of the alternatives in relation to the validity of the animal models. This is not a correct or fair legal test. The true test is whether alternative methods are sufficiently well developed and predictive of human responses to assure safety of human health to the same extent as animal models – or better. Alternative methods go through a lengthy validation process that includes an assessment of their predictivity. The results often are compared against the result of animal tests on the same chemicals, but attempts often are made to compare them against effects known in humans as well. For example, the reconstituted human skin epithelial models have been compared against human skin reactions and found to be more predictive than the rabbit test they now replace (Jirova et al., 2010). Assessing predictivity is difficult, however, because data on the gold standard (i.e., the human) is often limited, particularly for chemicals, since these, in general, should not be deliberately tested on humans. There are reviews of the predictivity of animal models for human effects, however, in which these are known from history of use. We present this information for each endpoint in order that it may be compared to similar data for the alternative methods. We do this on the premise that it is not possible to evaluate what the shortfalls of alternative methods are (and there are shortfalls) without first asking the question about what is or would be adequate.

3 The addition of skin sensitization and carcinogenicity to the 2013 endpoints

The experts were instructed by the Commission also to consider the endpoints of skin sensitization and carcinogenicity. While there is no issue with asking for an update on the progress in alternatives for these endpoints, there also is no legal basis for any legislative proposal to extend the deadlines with respect to these two endpoints. The possibility of extending the 2013 deadline mentioned in the text of the Cosmetic Directive only applies to repeated dose, toxicokinetics, and reproductive toxicity; neither skin sensitization nor carcinogenicity is listed in Article 18(2) of the recast Cosmetics Directive 1223/2009.

The Commission, in its 2004 report (SEC, 2004), which postdates the 7th Amendment agreed in 2003, made an assumption that the term "repeated dose" includes these endpoints, and they have continued to do so in subsequent reports. The Commission's argument seems to be that these tests can also be considered repeated-dose toxicity because animals may be subjected to more than one dose of the substance in question. This does not explain, however, why reproductive toxicity is listed separately, as tests for this endpoint also involve repeated dosing. It is our opinion that this position is untenable, since all these endpoints are terms of art, with a clearly recognized meaning in legislation, international guidelines, and toxicology industry usage. Carcinogenicity and skin sensitization always are listed as discrete endpoints, distinct from repeated-dose toxicity in EU legislation. Examples abound: in REACH (Regulation No 1907/2006) and the Test Methods Regulation (Regulation No 440/2008), the Pesticides Directive (Directive 91/414/EEC), the Biocides Directive (Directive 98/8/EC), the Medicines Directive (Directive 2001/83/EC), and the Veterinary Medicinal Products Directive (Directive 2001/82/EC). Even the SCCS in their notes of guidance (SCCP, 2006) and ECVAM reports on the status of alternative methods (Zuang et al., 2010) list these endpoints distinctly.

There is no written evidence from the time of the negotiations of the Cosmetics Directive testing and marketing bans that would suggest that the European Parliament intended that "repeated-dose" be used to cover several animal tests in the way the Commission is implying. Subsequent assumption on the part of the Commission does not alter the legal text. We have written to the Commission about this, and we intend to challenge if any proposal to extend the deadline also applies to these endpoints.

4 Repeated dose (chapter 1)

Criticisms of the experts' draft report

This chapter discussed the *in vitro* models for common targets of organ toxicity such as hepatotoxicity, nephrotoxicity, cardiovascular toxicity, neurotoxicity, and pulmonary toxicity, but it did not evaluate the evidence for the validity of these methods in isolation or combination. The experts concluded that the 2013 deadline could not be met by these methods, as there is a need to "reproduce integrated, whole-organism responses," and thereby fell into the high-fidelity fallacy trap. Although there was a section discussing the limitations of *in vivo* models, this was not quantitative, and no reference was made to studies looking at the predictivity (validity) of rodent models of sub-chronic effects. For example, the review of Olson et al. (2000), which found that rats and mice only predicted 43% of human effects for 150 pharmaceuticals, was not included, nor was the paper from Spanhaak et al. (2008), where concordance of hepatotoxic effects between rodents and humans was only 60% for 1,061 pharmaceutical compounds and 46% for another set of 137 compounds. Finally, it must be remembered that, although widely accepted, the procedure to derive Margin of Safety values from No Observed Adverse Effect Levels (NOAEL) in test animals is not validated for the purposes of predicting human health risks (Blaauboer and Andersen, 2007).

The chapter did not always give a neutral, complete, and quantitative evaluation of non-animal testing methods and their applicability to repeated dose end points. Actual developments were not reported, for example, on standardized organotypic lung models for sub-chronic or chronic toxicity testing (MucilAir™ and EpiAirway™), and new culture techniques that allow the maintenance of physiological functions over several weeks of co-cultured human hepatocyte (Schmelzer et al., 2009; Zeilinger et al., 2010), renal epithelial cell (Jennings, 2010), and primary cardiomyocytes (Sreeijt, 2008). In addition, industry strategies to include metabolism in *in vitro* models, as for

example, advanced new *in vitro* models to assess dermal penetration, including those of nanoparticles, and dermal metabolism (Jäckh, 2010; Landsiedel, 2010) could have been better detailed. Surprisingly, the report provided no information on the outcomes of Framework 7 project Predict-IV (on the optimization, standardization and characterization of the long-term human-based cell culture models utilized for assessing hepatotoxicity, nephrotoxicity and CNS toxicity) and Framework 6 project Predictomics (focused on the identification of biomarkers of chronic toxicity based on combined genomic, proteomic, and cytomic analysis of cells exposed to model hepatotoxins and nephrotoxins).

The use of Integrated Testing Strategies (ITS) to integrate *in vitro* models of various organ toxicities and *in silico* techniques was mentioned in the experts' report, but no specific strategies were discussed despite the chapter's conclusion that this is the way this endpoint may be replaced. No specific reference to the FRAME ITS (Grindon et al., 2008) was made in this context, and other strategies also were absent (Combes, et al., 2006; Prieto, et al., 2006; Boekelheide and Campion, 2010). As an aid to which organs need to be targeted by *in vitro* models, more information on the percentage of adverse effects seen across the organs could have been provided from data on human exposure to chemicals. The Boekelheide and Campion (2010) paper suggests a systematic approach to the analysis of results from batteries of *in vitro* tests, in analogy to a system of aircraft accident investigation. This new Toxicological Factors Analysis and Classification System can discriminate on a mechanistic level between different types of failures that are initiated by a toxicant. A manifest "active failure" as a last step is conditional on previous "latent failures." The system will allow the development of a fully fleshed out Taxonomy of Adverse Effects.

One admirable approach the authors of this chapter undertook was to ask companies what their strategies for avoiding animal testing were in relation to repeated dose. It was disappointing that only Unilever and Nestle responded to their request. These companies were employing the TTC approach in order to establish whether testing is genuinely necessary. The TTC approach is based on the concept that for all substances there is a level of exposure below which there is hardly any risk to human health, regardless of the toxicity of the substance. The level of exposure depends on very broad classes of likely toxicity; those chemicals not at all likely to be toxic can have higher exposure levels. With respect to cosmetics, ingredients such as preservatives, fragrances, and dyes are present in only tiny amounts within a product, and so it is possible that for many ingredients exposure will never exceed the TTC. Rather than new animal tests, then, all that is required is an evaluation, based on chemical structural similarity to other substances, as to the likely risk, which then allows a calculation of maximum daily exposure. This concept was used first for food additives, but COLIPA research has shown it to be relevant for dermal (Kroes et al., 2007) and inhalation (Westmoreland et al., 2010) exposure to cosmetics, and examples are now available; the SCCS is reviewing the concept's usefulness at the moment. The experts' report covered this approach but did not come to a conclusion about its usefulness or its range of applicability beyond the fact that it could "contribute to intelligent testing strategies to help reduce and refine animal use." In addition, other non-testing approaches for the risk assessment of long-term exposure to cosmetics were not adequately represented in the draft report, such as read-across or margin of safety values by grouping of chemicals, or weight-of-evidence considerations that take into account experience with previous consumer use (Weed, 2005).

Finally, the experts' report also suffered from "the common misconception that reliable QSAR models can be derived only for biological events with a common mode of action. It is important to remember that these methods do not model toxicological mechanisms but try to identify the relationship between compound properties and toxicological effects. With modern data mining and machine learning methods, reliable prediction models can be obtained from non-congeneric compounds, even for complex endpoints where many mechanisms may still be unknown. Models with improved predictivity and a broader applicability domain could be generated if software engineers would be granted access to the existing high quality test data which are not included in public databases" (Christoph Helma, personal communication).

An alternative analysis

Repeated dose information (NOAEL) is required for new cosmetic ingredients, but in many cases this can be avoided by use of the TTC concept, since substances are used in such low quantities that no adverse effects would be expected. In instances where this cannot be achieved, then a battery of *in vitro* tests should be employed, focusing on the liver which is the key target organ for repeated dose toxicity, followed by kidney, heart, nerves, lung, and immune system and selecting the more sensitive end point for the determination of the NOAEL (Prieto et al., 2006). Several *in vitro* models, developed as stand-alone methods, are at various development/validation stages in relation to most common targets for toxicity (Tab. 1). Although studies have shown these tests can predict effects seen in humans, the practical (but not insurmountable) problem remains: how to combine the results from several tests into a single "safety factor" for risk assessment purposes. Suggested approaches, such as in Prieto et al. (2006), could be used as a basis.

5 Skin sensitization (chapter 2)

Criticisms of the experts' draft report

Overall, we disagree with the author's conclusions that alternatives for risk assessment decision-making for skin sensitization are not yet available. This is because several *in vitro* methods show extremely high concordance with *in vivo* data, in the realm of 80% accuracy (e.g. 89% concordance of Direct Peptide Reactivity Assay with *in vivo* data on 82 chemicals), which is considered sufficient for ECVAM validation purposes (ECVAM, 2009). In addition, in contrast to the guinea

Tab. 1: An alternative approach for repeated dose

Alternative	Evidence of validity	Status
Step 1 – Low exposure substance *(no proteins, heavy metals, polyhalogenated-dibenzodioxins)* *No testing needed if human exposure is below 1800 mg/day for Cramer class III (low toxicity expected); 540 mg/day for Cramer class II (medium toxicity expected), 90 mg/day for Cramer class I (higher toxicity expected).*		
TTC	Relevance for cosmetics shown by COLIPA research (Kroes et al., 2007). Database on repeated dose oral toxicity data from 613 substances (Munro et al., 1996).	SCCS on-going evaluation for cosmetics.
Step 2 – Higher exposure substance *If human exposure exceeds TTC levels, conduct below in vitro testing in combination with QSARs for specific endpoints in a weight-of-evidence approach and select the more sensitive end point for determination of the NOAEL (Prieto et al., 2006).*		
Hepatotoxicity (liver)		
***In vitro* hepatotoxicity on human liver cell lines**	Pfizer study found 80% of 243 human hepatotoxicants were detected (O'Brien et al., 2006). 100% of 10 hepatotoxicants detected (Horii and Yamada, 2007).	Requires validation studies. Long term cell lines and cultures now available.
Nephrotoxicity (kidneys)		
***In vitro* kidney cell lines**	Good prediction of 15 nephrotoxicants *in vitro* (Duff et al., 2002).	ECVAM recommended validation studies in 1994 (Morin et al., 1997). On-going internal validation at Merck Serono (Hewitt, 2009).
Cardiotoxicity (heart)		
***In vitro* heart cells**	81% agreement between in vitro and clinical cardio-toxicity on 6 compounds (Schwengberg et al., 2004). Up to 97% agreement with *in vivo* test for 4 cardiotoxicants (Inoue et al., 2007).	Requires validation studies.
Neurotoxicity (nerves)		
***In vitro* neuronal cell test**	Excellent agreement with *in vivo* test for organophosphorus compounds (Malygin et al., 2003).	According to experts' report, ring trial ongoing with EU and US labs.
Pulmonary toxicity (lungs)		
***In vitro* lung epithelial cells** **EpiAirway™** **MucilAir™**	>81% correlation with existing human data with 11 chemicals on MucilAir™ (Huang et al., 2009).	Requires validation studies.
Immunotoxicity		
CFU-GM **(from bone marrow cells)**	Accurate prediction of *in vivo* results with 5 out of 6 test substances in pre-validation study. Positive results obtained on additional 20 substances (Pessina et al., 2001).	Validated by ECVAM in 2000 (ESAC statement, 2006).
***In vitro* human whole blood cytokine assay**	Results correlated well with the *in vivo* data on 31 compounds (Langezaal et al., 2002).	ECVAM pre-validation in 2002.
***In vitro* lymphocyte proliferation assay**	100% correct predictions on 6 chemicals (Carfi et al., 2007).	Progressing towards pre-validation (Lankveld et al., 2009).
Computer models for specific toxicity		
Computer models: **TOPKAT** **DEREK** **LAZAR**	TOPKAT (based on 393 chemicals from US EPA, FDA): able to predict 30% LOAELs within a factor of 3; 60% within a factor of 10; 96% within a factor of 100 (Tilaoui et al., 2007). LAZAR: 89% predictions within 1 log unit from experimental value (Maunz and Helma, 2008).	Accepted for regulatory purposes for cosmetics, biocides, plant protection products and chemicals (REACH).

pig maximization test (GPMT) or Buehler test, which only allows a crude estimate of potency (Keller et al., 2009), some *in vitro* assays provide information on sensitization potency (e.g. DPRA, hCLAT assays). Since several methods already have entered the prevalidation stage (e.g. DPRA, hCLAT and MUSST entered prevalidation in 2009), we disagreed with the timelines for replacement of this endpoint: "up to 2019" is an overly cautious timeframe. Under REACH Annex XI, methods can be used for positive prediction if they are suitable for entry into ECVAM pre-validation and for both negative and positive prediction if validated to internationally agreed protocols. It could be argued that the skin sensitization methods mentioned above would satisfy this already, and therefore we ask the question: if they are arguably suitable for predicting worker safety, why are they not (yet) suitable for predicting consumer safety of cosmetics?

Again, the experts' report fell into the high-fidelity fallacy trap by insisting that the complete mechanism of action of skin sensitization needed to be modeled on complete replacement. Not all experts agree with this perception. For example, Roberts and Patlewicz (2010) argue that haptenation (the reaction with protein) is the "single most important and possibly the only important step" in the prediction of skin sensitization. The extent to which a chemical will cause haptenation can be predicted by assessing its ability to react with proteins *in vitro*. Indeed, "whether a chemical is a sensitizer or not, and how potent it is if it is a sensitizer, depends on its chemical properties and on nothing else" (Dr. Dave Roberts, personal communication). The peptide reactivity tests have been criticized for not taking absorption and metabolism into account, but these experts contend that to do so would only underestimate the risk to the population at large (Roberts and Patlewicz, 2010).

In the *in silico* tools section, no reference was made to the work of Roberts and Aptula (2008) in the use of mechanistic domains within which simple and interpretable descriptors (logP and rate data) can be used to model the formation of the hapten and, in turn, skin sensitization. A mechanistically based paper that makes use of an *in silico* descriptor that is useful in modeling reactivity (and thus the LLNA) within the Michael domain is given under Enoch et al. (2008a). The same descriptor also has been used to model respiratory sensitization based on the same premise that haptenation is the key step that needs to be understood (the rest of the biology does not affect the sensitization outcome) (Enoch et al, 2010). The report also lacks a detailed discussion of a number of expert systems, such as QSAR model Toxtree, which can be used to predict potential skin sensitization mechanisms based on the Enoch encoding (Enoch et al., 2008b) or the Roberts rules for reaction mechanistic domains (Aptula and Roberts, 2006), and Derek for Windows, which has an extensive rule base able to identify skin sensitizers. The rule base within Derek for Windows is mechanistically based, taking the premise that haptenation is the key event that leads to skin sensitization. The use of multiple *in silico* tools can lead to weight of evidence approaches for the prediction of skin sensitization; a detailed discussion of such approaches is crucial. Finally, no reference was made to statistically-based models developed within the Framework 6 project CAESAR for skin sensitization and available for online use via the web (http://www.caesar-project.eu).These models have been developed and tested under stringent quality criteria to fulfill the principles laid down by the OECD, and the final models offer a robust and reliable method of assessing skin sensitization for regulatory use (Chaudhry et al., 2010).

In the section devoted to the animal test methods, such as the Local Lymph Node Assay (LLNA), GPMT or Buehler test, no evidence for their reliability, predictivity, or applicability was given. For example, it was not stated that, while the LLNA has been formally validated for hazard identification for regulatory purposes by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM, 1999) by comparing it with GPMT and available human data, such an evaluation has not been performed for the guinea pig tests. The ICCVAM review found that the LLNA or GPMT only predicted human reactions 72% of the time (SCCNFP, 2000). In addition, no details of the outcomes of key studies on non-animal methods are given – including predictivity and applicability domains, or on the ability of the *in vitro* assays to estimate potency. Therefore, no fair comparison with the current *in vivo* tests can be made. No reference is made to the TTC concept or, more specifically, to the Threshold of Sensitization Concern (TSC) concept recently proposed by Keller et al. (2009). The TSC values (0.91 or 0.30 μg/cm^2 dependent on chemical class) derived in terms of amount per skin area, based on human skin sensitization data on 53 fragrance ingredients from the Research Institute for Fragrance Material of the International Fragrance Association dataset, largely support the dermal sensitization thresholds. Finally, there was no update from the Framework project Sens-it-iv and no section devoted to ITS for skin sensitization developed under the Framework project OSIRIS.

Our proposed approach

Our suggestion is that the TSC approach may be used to mitigate any testing if the ingredient is used in very small quantities. If this approach cannot be used, then a combination of QSAR and *in vitro* peptide reactivity tests may be sufficiently predictive. The mechanism of how skin reacts to "sensitizing" substances is actually well understood and "haptenation", the reaction of proteins in the skin to the substance, is considered the key step. It is therefore possible to determine the skin sensitization potency of a substance based on how it binds to proteins. The Direct Peptide Reactivity Assay, DPRA, used by industry since the early 2000's, has almost completed ECVAM pre-validation. Evidence already indicates that this test alone can predict 89% of substances and that further development of a model to consider metabolism would only underestimate the risk to humans. Computer models alone also have similar predicting strength. In addition, two *in vitro* methods using skin cells (MUSST and h-CLAT) are being pre-validated by ECVAM with results due in 2011. How these can be used in a strategy is illustrated in Table 2.

Tab. 2: An alternative approach for skin sensitization

Alternative	Evidence of validity	Status
Step 1 – Low exposure substance, no very strong sensitizers *No testing needed if human exposure below 0.91 or 0.30 µg/cm² skin area dependent on chemical class (Keller et al., 2009).*		
Threshold of Sensitisation Concern (TSC) concept	Applicability to skin sensitization evidenced with a meta-analysis on human data from 53 fragrance allergens (Keller et al., 2009). Main application with rinse-off products (e.g. shampoos, soap) and low concentration chemicals in stay-in (e.g. hair spray, gel) or leave-on products (e.g. face cream) (Keller et al., 2009).	SCCS on-going evaluation for cosmetics.
Step 2 – Higher exposure substance *If human exposure exceeds TSC levels, run a QSAR. If negative, follow up with DPRA. In case of doubt conduct MUSST or hCLAT.*		
QSAR computer models	83% correct classifications for DEREK, 73% for TOPKAT (Fedorowicz et al., 2008). TOPS-MODE used to screen 229 hair dyes (Søsted et al., 2004). CAESAR made 90% correct predictions on 42 chemicals (Chaundry et al., 2010). OECD Toolbox has data on 600-800 substances for "read across".	Accepted for regulatory purposes for REACH; each model has to be validated according to OECD principles.
DPRA (Direct Peptide Reactivity Assay)	94% agreement with *in vivo* data on 18 chemicals (Ahlfors et al., 2003). 89% agreement with *in vivo* data on 82 chemicals (Gerberick et al., 2007).	ECVAM pre-validation on-going (results expected 2011). Improved sensitivity for weak sensitizers (Natsch et al., 2007).
MUSST (Myeloid U937 Skin Sensitization Test)	93% of 16 chemicals correctly predicted in Proctor and Gamble study (Python et al., 2007). Used by L'Oréal on more than 800 chemicals, ok for 80% of them (Martinozzi, 2010).	ECVAM pre-validation on-going (results expected 2011). Improvements for water insoluble, coloured, toxic substances; metabolic capabilities have been included.
hCLAT (human Cell Line Activation Test)	Evaluated by 5 labs (P&G, Shiseido, Kao, Henkel and L'Oreal) since 2004 (Anon, 2008). Studies at Shiseido show 93% correct predictions in 29 chemicals (Sakaguchi et al., 2009) and 84% agreement in 100 chemicals (Ashikaga et al., 2010).	JaCVAM (lead) – ECVAM pre-validation study on-going (results expected 2011).

6 Carcinogenicity (chapter 3)

Criticisms of the experts' draft report

While this chapter provided an honest assessment of the current requirements for carcinogenicity testing based on the SCCS Information Requirements (SCCP, 2006), it also fell into the high fidelity fallacy trap and did not provide an adequate assessment of the validity and reliability of either *in vivo* or *in vitro* methods.

As stated by experts, the two-year cancer bioassay is rarely conducted as it is costly, lengthy, and has animal welfare implications. The SCCS do not require carcinogenicity tests unless "considerable oral intake or dermal absorption is expected." This is confirmed by Pauwels et al. (2009) who showed that carcinogenicity data were seen in less than 40% of submissions to the SCCS between 2000 and 2006. Given that these submissions are for cosmetic ingredients of particular concern (dyes, preservatives, and UV filters) one might expect that the prevalence of carcinogenicity data among other cosmetic ingredient dossiers is even lower. The experts rightly point out that the current strategy for carcinogenicity is to assess for likely genotoxicity using *in vitro* methods. Ingredients that are positive in these assays are not taken forward for development. Added confidence in a substance's lack of carcinogenic potential is then provided by conducting a repeated dose test. Thus, the impact of a ban on carcinogenicity tests per se is minimal. It was therefore both disappointing and quite surprising to see the experts proceed to warn against abolishing *in vivo* carcinogenicity tests. What should have been attempted, in our opinion, is an assessment of whether *in vitro* tests and other approaches, such as TTC, could provide not only adequate safety factors but also decrease the impact on the development of new ingredients.

According to Kirkland et al. (2005), 93% of 553 rodent carcinogens were detected in at least one of the three most common *in vitro* genotoxicity tests (Ames-Test, Mouse Lymphoma Assay, and *in vitro* Micronucleus Test or Chromosomal Aberrations Test). Combinations of two and three test systems had greater sensitivity than individual tests, resulting in sensitivi-

ties of around 90% or more, depending on test combination. The specificity of the Ames test was reasonable (73.9%), but all mammalian cell tests had very low specificity (i.e. below 45%), and this declined to extremely low levels in combinations of two and three test systems. The experts highlighted the impact of the risk of detecting false positives with the *in vitro* genotoxicity assays, using the example of a review of hair dyes from Speit (2009). In this review, the sole use of *in vitro* tests may have resulted in a number of false positives and therefore withdrawal from the market of these products. The application of the over-protective criteria may not be a limitation, however. Indeed, protecting the public is far more important than producing new hair dyes. In the absence of human data, it is indeed possible that these false positives are not "false" at all. It should also be noted that a significant proportion of the hair dyes were deemed safe in this assessment. Nonetheless, the report authors did not appreciate the impact of a new strategy to reduce the percentage of false positive *in vitro* genotoxicity tests, thus increasing test predictivity, with respect to the need for *in vivo* genotoxicity/carcinogenicity testing (Fowler et al., in press; Kirkland and Fowler, 2010; Parry et al., 2010).

No data on validity and predictivity of *in vivo* tests is given, and therefore a neutral evaluation cannot be conducted. According to Ennever et al. (1987), the sensitivity of animal bioassays is very high (all definite human carcinogens adequately tested were positive). The specificity is low, however (in 20 of the probable non-carcinogens tested for rodent carcinogenicity in animal bioassays, 19 were positive and only one was negative). Little attempt has been made to validate the lifetime rodent bioassay against human carcinogenicity (Ennever and Lave, 2003). A survey of the US Environmental Protection Agency database to assess the human utility of animal carcinogenicity data showed the animal data were predictive for 42% of chemicals. For the 128 chemicals with human or animal data assessed, however, human carcinogenicity classifications were similar only for those 17 possessing significant human data. The authors concluded that the problem with animal carcinogenicity tests is not their lack of sensitivity for human carcinogens, but rather their lack of human specificity (Knight et al., 2005).

A retrospective analysis conducted using the National Toxicology Program database on sixteen chemicals that may lead to liver, lung, or kidney tumors in two-year rodent cancer bioassays – and for which short-term data also were available – showed that cancer often is secondary to a biological precursor effect, the mode of action sometimes is not relevant to humans, and key events leading to cancer in rodents from nongenotoxic agents usually occur well before tumorigenesis and at the same or lower doses than those producing tumors (Boobis et al., 2009).

The authors concluded that the two-year bioassay in rats and mice is, at best, only an indicator of potential hazard. Similar conclusions were reached by Ward (2007), who observed that rodents do not commonly develop the spontaneous tumors most prevalent in humans, including those of the colon and prostate. This is due, in part, to differences in genetics, diet, specific natural chemical exposure, and infectious agents.

In addition to the genotoxicity assays that already have been validated and received regulatory approval, the cell transformation assays (CTA) seem to offer the most promising replacement options. The experts did not provide details on the predictivity and reliability of these tests, however, and therefore appeared too hasty in their dismissal of the opportunities the tests may provide for complete replacement. Data on rodent and human predictivity of the Syrian Hamster Embryo (SHE) assay published by Long (2007) showed that SHE has a concordance with the rodent bioassay ranging from 85% (SHE pH ≥7) to 74% (SHE pH 6.7), sensitivity 92% (SHE pH ≥7), specificity 85% (SHE pH 6.7) and predictivity 88%. A meta-analysis done by the OECD indicated that the three CTA assays have an overall sensitivity of 90% of class I (known) and 95% of class II (possible/probable) human carcinogens (OECD, 2007). In comparison to this, the rodent bioassay was calculated to have a sensitivity of 50% or 90% on human carcinogens, depending on how the results are interpreted (Ennever and Lave, 2003). The SHE (both pH ≥7 and pH 6.7) correctly identified 100% of the 44 inorganic human carcinogens tested and was able to identify 9 out of 11 organic carcinogens – a sensitivity of 82% (OECD, 2007). "The limitations to these tests seem minimal, provided that cell clones that retain enough metabolizing capacities to detect different classes of chemicals acting as genotoxic compounds through the formation of stable adducts to DNA are used and more than one *in vitro* test is performed to improve reliability and predictability, the complete replacement looks like a real possibility" (Annamaria Colacci, personal communication).

In the QSAR section, we suggested additional work by Fjodorova et al. (2010) under the EU Framework CAESAR project and two QSAR models for carcinogenicity developed by Contrera et al. (2007). The section on TTC was well developed, although it was almost forgotten in the Conclusions, which stated that when repeated dose toxicity is banned, methods for quantitative detection of non-genotoxic carcinogens will be *limited* to tools such as read across, QSAR, and TTC. "It should be better explained that the use of TTC and read across are not a limitation for *safety* but for the development of new cosmetic ingredients. The NOAEL and the application of the safety (better to say uncertainty) factor is not always considered the best approach to protect human health, and it must be remembered that the "safety" factor is an arbitrary number that is applied to take into account interspecies and intraspecies differences when extrapolating from animal studies, thus it is not the panacea" (Annamaria Colacci, personal communication.)

Our proposed approach

We propose that genotoxic carcinogens can be identified by a number of long-standing *in vitro* cell based tests. These tests allegedly have been over-sensitive, but newer tests are more predictive. A more complete method based on CTA also has been in use for more than 40 years but only recently entered an ECVAM pre-validation study. Experts agree that carcinogenicity studies are rarely conducted, as they are expensive and time-consuming, are not specified under the Cosmetic Directive, and are rarely requested by the SCCS. A combination of the accepted *in vitro* genotoxicity tests, the CTA assay, and exposure-based TTC approaches (providing a precautionary approach for consumers) should be the preferred approach, see Table 3.

Tab. 3: An alternative approach for carcinogenicity

Alternative	Evidence of validity	Status
Step 1 – Low exposure substance, no high potency carcinogen (aflatoxin-like, azoxy and N-nitroso compounds) *No testing needed if human exposure below 1.5 µg/day for chemicals with no structural alerts for genotoxicity and 0.15 µg/day for chemicals with structural alerts for genotoxicity.*		
TTC	Values derived from Carcinogen Potency Database (CPDB) including data on more than 700 chemical carcinogens (Kroes et al., 2004). Proposed use with genotoxic impurities in drugs (Bercu et al., 2010).	SCCS on-going evaluation for cosmetics.
Step 2 – Higher exposure substance *If human exposure exceeds TTC levels, perform an Ames test and one of the other in vitro genotoxicity tests. If both are positive assume genotoxic carcinogen; in case of doubt conduct the CTA assay.*		
Genotoxic carcinogens		
Bacterial reverse mutation assay (Ames test)	Developed in the late 1950s. Well established and scientifically accepted test. 90% rodent carcinogens detected when combined with MLA and MNT assays (Kirkland et al., 2005). 77% accuracy on 368 chemicals (Zeiger, 1998).	Accepted for regulatory purposes (OECD TG 471, 1997).
***In vitro* gene mutation assay in mammalian cells (MLA)**	90% of 553 rodent carcinogens detected when combined with MNT and Ames test (Kirkland et al., 2005).	Accepted for regulatory purposes (OECD TG 476, 1997).
***In vitro* Chromosome Aberration assay in mammalian cells (CA)**	85% of 553 rodent carcinogens detected when combined with Ames test and MLA (Kirkland et al., 2005).	Accepted for regulatory purposes (OECD TG 473, 1997).
***In vitro* Micronucleus Test in mammalian cells (MNT)**	90% of 553 rodent carcinogens detected when combined with MLA and Ames test (Kirkland et al., 2005). 83% agreement on 113 chemicals in ECVAM validation study (Corvi et al., 2008).	Validated by ECVAM 2006 (ESAC, 2006) Accepted for regulatory purposes (OECD TG 487, 2010).
Genotoxic and non-genotoxic carcinogens		
Cell Transformation Assays (CTA with SHE, Balb/3T3 and Bhas 42 cells)	Assays established since late 1960s. OECD review in 2007 concluded that 90-95% human carcinogens could be detected (OECD, 2007). ECVAM workshop found that 80-83% rodent carcinogens were detected with 213 chemicals (Combes et al., 1999). Proctor and Gamble study showed 85% agreement with rodent data with 56 chemicals (LeBoeuf et al., 1996). Pfizer study showed 89% agreement with rodent data with 19 chemicals (Mauthe et al., 2001).	Development of test guideline recommended by OECD in 2006 (OECD, 2007). ECVAM pre-validation completed in 2009 for SHE and Balb/3T3, ongoing for Bhas 42 (statement was expected in 2010).

7 Toxicokinetics (chapter 4)

Criticisms of the experts' draft report

Overall, we considered this chapter the most comprehensive and balanced, with the experts approaching the problem with the assumption that animal testing would be banned in 2013, their so called "2013 non-animal approach scenario." This enabled the experts to be more creative in their analysis of how toxicokinetics could be studied, given this scenario. The experts concluded that the approach to toxicokinetics from an *in vitro* or *in silico* basis was "well understood" and that "a whole array of *in vitro* and *in silico* methods at various levels of development is available for most of the steps and mechanisms that govern the toxicokinetics of cosmetic substances." They expressed concern that renal models are less well developed, so we referred them to the repeated dose chapter for evidence of more extensively developed models.

The positive and thorough assessment of the potential for replacing the toxicokinetics endpoint was not reflected, however, in the timelines given by the experts, which seemed overly conservative given the description of the status of these methods in the text. In addition, the report did not give enough emphasis to the fact that the updated OECD Test Guideline 417 on toxicokinetics already foresees the use of *in vitro* studies with microsomal fractions to address metabolism or the potential for induction of biotransformation, the use of *in vitro* dermal absorption studies to characterize absorption, and the use of toxicokinetic modelling for the prediction of systemic exposure and internal tissue dose.

The experts explain that, in fact, toxicokinetics is rarely a requirement under any safety legislation but is considered a

"nice to have" endpoint, a matter over which they expressed regret. The point, nonetheless, is that, toxicokinetic information, although useful, is not specified in the Cosmetic Directive and is not considered a "core requirement" by the SCCS (SCCP, 2006). Not surprisingly, therefore, the survey of dossiers submitted to the SCCS between 2000 and 2006 found that fewer than 50% of dossiers included *in vivo* toxicokinetic data (Pauwels et al., 2009). It appears that *in vitro* methods (skin absorption) and the use of *in silico* physiologically-based pharmacokinetics (PBPK) models is already commonplace, and so it was disappointing not to see examples of current (as opposed to future) company strategies for toxicokinetics, as attempted in the repeated dose chapter.

We also found a few areas where additional information on the utility and validity of models could be found. For example, according to published evidence, the suitability of the artificial PAMPA-skin for skin absorption (Ottaviani et al., 2007) and of the *in vitro* HaCat cell model (Goebel et al., 2009) is already partially established for cosmetics, thus the estimated timeline to enter pre-validation should have been sooner than 2013. The same timeline update was suggested for metabolic activation, for which the Ames test is available. Finally, perhaps due in part to its forward- looking nature, the report did not evaluate current *in vivo* methods, which, in our opinion, are limited due to the significant differences in metabolism and physiology between animals and humans. For example, before *in vitro* Absorption, Distribution, Metabolism and Excretion studies (ADME) on human cell models were routinely used by the pharmaceutical industry, the failure rate of drugs in clinical trials due to poor prediction of ADME was 40% (Kola and Landis, 2004); now it is only 10% (McKim, 2010).

Our proposed approach

As described by the experts, relevant stages of toxicokinetics can be modeled using mathematical Physiologically Based ToxicoKinetic models (PBTK). These models consist of a set of physiological and chemical parameters that can predict the distribution and excretion of substances through the human body following initial input of information on absorption and metabolism. The pharmaceutical industry has used these tests with growing sophistication since the 1970s (Andersen, 2003). The skin is the main route for the absorption of cosmetics and can be modeled using the regulatory approved *in vitro* skin method. Metabolism can be predicted through the use of high-

Tab. 4: An alternative approach for toxicokinetics

Alternative	Evidence of validity	Status
Step 1 – Determine likely absorption *Conduct a skin absorption assay and use together with physicochemical properties to determine likely systemic absorption through the skin or other routes.*		
***In vitro* dermal absorption test**	*In vitro-in vivo* correlation evidenced since early 1980s (Bronaugh et al., 1982). OECD experts agreed in 1999 that there was sufficient data to support the Test Guideline (OECD, 2004). Validation study on new reconstituted human epithelial models demonstrated appropriateness on 8 OECD test chemicals (Schäfer-Korting et al., 2008).	Basic criteria for the use for cosmetics first published by SCCNFP (now SCCS) in 1999 (SCCNFP, 1999). Accepted for regulatory purposes (OECD TG 428, 2004).
Step 2 – Determine Distribution, Metabolism and Excretion *If absorption possible, use input from absorption assay to model using a combination of PBPK models and in vitro assays.*		
PBTK computer models	80% correct *in vivo* predictions of distribution for 123 drugs within 2-fold error (Poulin and Theil, 2002). 70% of 19 human drugs would have been predicted for pharmacokinetics by PBPK models alone (Jones et al., 2006). 90% correct predictions of renal excretion for 40 compounds (Manga et al., 2003). 88% precision of predicted renal clearance for 141 drugs (Kusama et al., 2010).	Use proposed by EFSA for pesticide residues in food (EFSA, 2007). Use included in regulatory guidelines (OECD TG 417, 2010). ECVAM workshop in 2007 set guidelines for their use (Bouvier d'Yvoire et al., 2007).
Metabolism and Excretion		
***In vitro* assays on hepatocytes (liver cells)**	Review of studies concluded that hepatic clearance could be predicted using human liver microsomes (Chiba et al., 2009). Retrospective analysis on 50 drugs found human liver cells are as predictive as animal tests (Hosea et al., 2009). *In vitro* tests with PBPK modelling (SCHH-PBPK) gave better prediction accuracy for humans compared to *in vivo* rat and dog (Yamazaki et al., 2011).	Being pre-validated by ECVAM in 2011. Included in regulatory guideline (OECD TG 417, 2010).

throughput assays on cultured human hepatocytes, which are commonly used in most pharmaceutical companies. A proposed approach, summarized in Table 4, would not provide a complete ADME analysis but may provide adequate data to help make safety decisions.

8 Reproductive toxicity (chapter 5)

Criticisms of the experts' draft report

This chapter was one of the weaker chapters in that it did not justify many of its conclusions on the utility of the methods and, inexplicably, it omitted from its table of methods *in vitro* and QSAR models that had been discussed. While the chapter was clear in its assessment of the need for reproductive toxicity testing for cosmetics (tests are only required if "considerable oral intake or dermal absorption is expected" (SCCP, 2006)), it was overly negative regarding the value of existing *in vitro* methods and unimaginative in its approach to the 2013 deadline. The experts in this chapter failed to consider the TTC approach and did not quantitatively assess the validity of *in vivo* or *in vitro* or *in silico* methods. Much of the chapter was devoted to describing qualitatively the various *in vivo* methods available, ignoring the fact that for cosmetics ingredients only the developmental toxicity guideline (OECD TG 414) tends to be used (Rogiers and Pauwels, 2008).

No evidence of the validity of the *in vivo* test methods was given. This is a concern because there is evidence that the predictive power of the prenatal developmental toxicity test is rather poor. For example, Hurtt et al. (2003) found that the positive predictivity of one species to teratogenic effects in rat, mouse, or rabbit was around 60% for 105 veterinary pharmaceuticals. Bailey et al. (2005) found that the rat was positively predictive of 35 known human teratogens in 61% of cases and the rabbit in 41%.

Failure to assess the *in vivo* methods is also a concern because it does not allow a fair comparison with potential *in vitro* or *in silico* methods, which may have similar or better predictivity. No data on the predictive capacity of the Whole Embryo Culture (WEC) test (Genschow et al., 2002), the micromass test (MM) (Spielmann et al., 2004), or the zebrafish embryo test (Selderslaghs et al., 2009) was given, even though it is easily available in the references given here. Perhaps more crucial, no information is given on the validation outcomes of the embryonic stem cell test (EST), a more commonly used, more refined test. Indeed, the test was omitted from the table of methods used by all chapters to provide an overview of all methods. The ECVAM validation in 2002 stated that "the correlation between the *in vitro* data and *in vivo* data was good (accuracy 78%) and the test proved applicable to testing a diverse group of chemicals of different embryotoxic potentials" (ESAC, 2002). The expert report stated that the EST method is limited but did not provide details relating to this comment nor to the counterpoint to this criticism. For example, both Spielmann (2009) and Combes (2009) criticize the ReProTect study that appeared to demonstrate a weakness of the test because it changed the classification from that used in the original ECVAM validation study. The EST has now been improved to increase both its applicability (Dartel, et al., 2010) and its speed of conduct (Peters et al., 2008a,b) and to account for metabolism (Hettwer et al., 2010). This was not considered by the experts. No references of industry use of the EST were given. For example, Pfizer uses the EST to make compound development decisions and Johnson and Johnson also have developed an automated system for HTP of the EST (Peters et al., 2008a,b). Most recently, West et al. (2010) found that the model was able to correctly predict the teratogenicity of seven out of eight blinded drug treatments, with a specificity of 100%, sensitivity of 80% and overall accuracy of 88%.

The experts' coverage of more complex assays, such as the endocrine transcriptional assays, was confused and incomplete. Some methods were featured in the table of methods but were not referred to within the text, and some methods that were referred to in the text did not appear in the table. This is of particular concern when some of these methods are considered validated or in the process of being validated – for example, the estrogen receptor transcriptional assay (LUMICELL-ER) and the Stably Transfected Transcriptional Activation assay (STTA), now OECD TG 455.

In conclusion, given that several receptor binding assays are in draft form or validated at OECD, there are some valid QSAR models, and several *in vitro* assays have been validated by ECVAM, it would have been appropriate to present a draft testing strategy similar to one we offer in Table 5. Indeed, this is what the ReProTect study has just done in its "feasibility study" (Schenk et al., 2010). We think the experts should have considered the outcomes of the ReProTect feasibility study and used these as a basis for discussion of next steps. We are not alone in this opinion; Dr Spielmann has made the same point (Spielmann, 2010). At the very least, the expert report could have illustrated the key stages of the reproductive cycle that should be covered and suggested where the gaps in available methods are, whether this is in applicability domain, predictability, or coverage of predictive endpoints. It may not be assumed that all aspects of the reproductive cycle are (equally) crucial to be represented (by alternative methods). For example, Bremer (2008) stated that "I must insist that it is embryonic development, rather than fetal development, which is the principle cause for concern, since organogenesis is the most sensitive phase in the developing child."

Our proposed approach

Several methods, including whole embryo cultures, stem cell tests, and receptor binding assays have been developed and are either validated according to ECVAM principles and/or are already OECD guidelines. We argue that, individually, some of these methods already show sufficient predictability of human effects across a range of test chemicals, see Table 5. It may not be necessary to cover all stages of the reproductive cycle, as some are more sensitive to chemicals than others. For example, the EST covers the development of the embryo, which is a very sensitive period. Thus a combination of these methods, covering the most sensitive endpoints in the reproductive cycle, may already be able to predict reproductive toxicity to an acceptable level of certainty. The EU ReProTect project recently concluded

Tab. 5: An alternative approach for reproductive toxicity

Alternative	Evidence of validity	Status
Step 1 – Low exposure substance *No fertility testing if exposure below 1.5 µg/kg bw/day (oral), 1.0 µg/m³ (inhalation). No developmental toxicity testing if exposure below 1.0 µg/kg bw/day (oral), 0.5 µg/m³ (inhalation) (Bernauer et al., 2008).*		
TTC	Values derived from fertility and developmental toxicity data (oral and inhalation exposure) on 91 chemicals (Bernauer et al., 2008).	SCCS on-going evaluation for cosmetics.
Step 2 – Higher exposure substance *If human exposure exceeds TTC levels, perform a combination of the validated in vitro embryotoxicity tests.*		
Embryonic Development		
***Ex vivo* rodent Whole Embryo Culture test (WEC) Micromass Test (MM)**	Widely used by industry for screening for developmental toxicants. ECVAM validation study: Up to 80% accuracy with 14 chemicals (100% accuracy with strong embryotoxicants (Genschow et al., 2002).	Validated by ECVAM in 2002 (ESAC statement, 2002).
(mouse/human) Embryonic Stem Cell Test (EST)	Widely used by industry for screening for developmental toxicants. ECVAM validation study: 78% agreement for 14 chemicals (100% for strong embryotoxic chemicals) (Genschow et al., 2002). 75% agreement with *in vivo* results for 63 chemicals (Paquette et al., 2008). 88% accuracy for 8 drugs (West et al., 2010).	Validated by ECVAM in 2002 (ESAC statement, 2002). Improvements have been recently made to increase applicability (Dartel et al., 2010), speed of the assay (Peters et al., 2008) and to account for metabolism (Hettwer et al., 2010).
Step 3 – Higher exposure substance, non-embryotoxic *If human exposure exceeds TTC levels, and the substance is non-embryotoxic, perform a combination of fertility and endocrine in vitro tests to determine likely effects on fertility.*		
Male fertility		
Computer Assisted Sperm Analysis (CASA)	Test evaluated by two different laboratories in more than 35 chemicals (AXLR8, 2010).	Pre-validated in ReProTect project.
Testicular fragment culture	82% expected results on 11 chemicals (Freyberger et al., 2010b).	Needs to be taken forward for prevalidation.
Leydig cell test	"Good" results on 15 chemicals (AXLR8, 2010). Detected all 5 endocrine disruptors (La Sala et al., 2010).	Needs to be taken forward for prevalidation.
Sertoli cell test	"Good" results in two laboratories for seven chemicals (AXLR8, 2010).	Needs to be taken forward for prevalidation.
Female fertility		
bovine *in vitro* (oocyte) maturation, bIVM	Good correlation with *in vivo* effects for 15 chemicals (Lazzari et al., 2008), good inter-laboratory variability on 8 chemicals (Luciano et al., 2010).	Pre-validated in ReProTect project.
Endocrine Effects		
Estrogen receptor alpha binding assay	Bayer Schering study showed it reliably ranked compounds with strong, weak, and no effect with high accuracy on 12 chemicals (Freyberger et al., 2010a).	Part of OECD/ReProTect project, expected to go to ECVAM validation.
Estrogen Receptor (ER) – Transcriptional Activation Assay, MELN	Bayer Schering pre-validation study showed good accuracy on 16 chemicals and good inter laboratory consistency (Witters et al., 2010).	ECVAM pre validation report due 2011, expected to go to ECVAM validation.
AR CALUX reporter gene assay	Inter-laboratory study on 64 chemicals showed 74% agreement (Sonneveld et al., 2006). Pre-validation study showed excellent agreement for 14 out of 16 chemicals (Van der Burg et al., 2010). Up to 85% agreement with rabbit test for 50 chemicals (Sonneveld et al., 2011).	Pre-validated in ReProTect (AXLR8), expected to go to ECVAM validation.

Alternative	Evidence of validity	Status
Estrogen receptor transcriptional assay, LUMICELL-ER	All 28 estrogen disruptors were detected (Gordon and Clark, 2005).	ICCVAM validation report expected 2011.
Stably Transfected Transcriptional activation assay (STTA) estrogen	80% accuracy on 46 chemicals (CERI, 2006).	Validated by CERI in 2006. Accepted for regulatory purposes (OECD TG 455, 2009).
H295R Steroidogenesis assays based on a human cell line	78% accuracy for testosterone effect on 18 chemicals, 88% for estradiol effect on 16 chemicals (OECD, 2009). "Overall, these results indicate that…the H295R would always flag a chemical as a potential disruptor of steroidogenic processes or a reproductive toxicant" (OECD, 2009).	Validated by OECD/EPA in 2009. Draft OECD test guideline being discussed.

that a battery of cell tests "allowed a robust prediction of adverse effects on fertility and embryonic development" (Schenk et al., 2010), with a combined accuracy of between 70-100% for 10 test chemicals (AXLR8, 2010). The use of these tests should be viewed in the context of the poor predictivity of the animal test and the fact that these tests are not always considered mandatory – due, in part, to the low exposure of humans to individual cosmetic ingredients. Those companies that voluntarily undertake reproductive toxicity tests usually only do the developmental toxicity test (Rogiers and Pauwels, 2008), which the EST may effectively replace. In addition, the threshold of toxicological concern (TTC) approach has demonstrated feasibility for reproduction end points for chemicals generally (Bernauer et al., 2008) and may also be used in certain cases when exposure is low.

9 Conclusion

In general, most chapters took the approach that the animal test has to be mimicked in full, a common, yet arguably, incorrect assumption called the "high fidelity fallacy" (Russell and Burch, 1959). Not surprisingly, the experts believed it would take a very long time to mimic the animal test completely. We disagree that completeness is more important than or as important as predictivity, and we would have liked to see a thorough assessment of this. While we were pleased to see a consistent approach to the presentation of all methods through the use of a table, this did not include a quantitative assessment of their suitability and therefore made it easier for some methods to be dismissed without apparent evaluation of the data on their predictive capacity. Examples of highly predictive tests that were not considered effective replacements because of this included the embryonic stem cell test for developmental reproductive toxicity, the peptide reactivity tests for skin sensitization, and the cell transformation assays for carcinogenicity.

The reports also were very inconsistent in their approach to the problem, their evaluation of the need for animal testing, the reliability of animal tests, the status of alternatives, and other approaches to waive animal testing, such as exposure-based techniques. Not all reports covered the possibility of waiving animal tests using the TTC approach, which is an approach used for substances that are applied in very small quantities. Coverage of other approaches such as QSARs and ITS also varied between the chapters, with the chapter on repeated dose covering these but not the chapter on reproductive toxicity. In contrast to the other chapters, the chapter on toxicokinetics started from the basis that animal testing was no longer permissible, and a more forward-looking assessment was the result. Similarly, the chapter on repeated dose included company strategies, which was helpful in identifying what is being used in industry practically as opposed to theoretically. It showed that companies already are applying imaginative approaches to avoid testing on animals and placing potentially harmful ingredients on the market.

Our paper summarizes these comments but also provides the kind of information and approach that we expected to see. The alternative approaches for each endpoint presented here are not meant to be complete answers to the problem but rather to provide a good basis for discussion about the immediate utility of various methods, if not to provide reassurance that some methods soon will be a complete answer. It would have been useful if the experts had started from this point and then – as was requested of them – highlight any areas that are not yet adequate and provide timelines for when they might be. In most cases, the experts failed completely to provide reasonable timescales for a new deadline for testing or a strategy for the steps that need to be made to ensure that alternatives are available. Some reports failed to suggest possible deadlines at all.

Finally, the possibility of extending the 2013 deadline in the text of the Cosmetic Directive only applies to Repeated Dose, Toxicokinetics, and Reproductive Toxicity. There has been a subsequent assumption over the years on the part of the Commission that the term "repeated dose" also includes skin sensitization and carcinogenicity endpoints, which we believe is incorrect. The evidence that replacements are nearly validated for these endpoints should help convince legislators that any extension to the 2013 deadline should not be applied to these tests.

In conclusion, the reports were not a complete evaluation of the status of alternative methods, as they are not detailed enough and are inconsistent in their approaches. The fundamental problem, albeit a commonly held one, is an assumption that all aspects of the *in vivo* approach would need to be modeled in order to provide complete replacements for these endpoints. This inappropriate assumption has naturally led the experts to an "easy" conclusion – i.e. that the 2013 deadline will not be met. This has been a missed opportunity to review the status of alternatives thoroughly, discuss the genuine obstacles objectively, and provide a workable framework for replacement. Other experts have criticized the report along similar lines (see Balls and Clothier, 2010; Spielmann, 2010). As they did, we recommend that the Commission not pay much heed to this report unless or until substantial amendments are made in the final version. We look forward to reading the final report.

References

Aeby, P., Ashikaga, T., Bessou-Touya, S. et al. (2010). Identifying and characterizing chemical skin sensitizers without animal testing: Colipa's research and method development program. *Toxicol. In Vitro 24*, 1465-1473.

Ahlfors, S. R., Sterner, O. and Hansson, C. (2003). Reactivity of contact allergenic haptens to amino acid residues in a model carrier peptide, and characterization of formed peptide-hapten adducts. *Skin Pharmacol. Appl. Skin Physiol. 16*, 59-68.

Andersen, M. E. (2003). Toxicokinetic modeling and its applications in chemical risk assessment. *Toxicol. Lett. 138*, 9-27.

Anon (2008). EU FP7 Project Sens-it-iv Newsletter 13. http://www.sens-it-iv.eu/files/newsletter/Sens-it-iv_Newsletter_13.html

Anon (2010). Terms of reference "Report on alternative (non-animal) methods for cosmetics testing: Current status and future prospects – 2010". European Commission, Health and Consumers Directorate-General (DG SANCO), document, no longer on the website.

Aptula, A. O. and Roberts, D. W. (2006). Mechanistic applicability domains for non animal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol. 19*, 1097-1105.

Ashikaga, T., Sakaguchi, H., Sono, S. et al. (2010). A comparative evaluation of in vitro skin sensitisation tests: The human Cell-line Activation Test (h-CLAT) versus the Local Lymph Node Assay (LLNA). *ATLA 38*, 275-284.

AXLR8 (2010). Alternative Testing Strategies Progress Report 2010. AXLR8 Consortium Report, http://www.AXLR8.eu

Bailey, J., Knight, A. and Balcombe, J. (2005). The future of teratology research is in vitro. *Biogenic Amines 19*, 97-145.

Bailey, J. and Taylor, K. (2009). Comment: The SCHER report on non-human primate research – biased and deeply flawed. *ATLA 37*, 427-435.

Balls, M. and Clothier, R. (2010). A FRAME response to the Draft Report on Alternative (Non-animal) Methods for Cosmetics Testing: Current Status and Future Prospects–2010. *ATLA 38*, 345-353.

Bercu, J. P., Morton, S. M., Deahl, J. T. et al. (2010). In silico approaches to predicting cancer potency for risk assessment of genotoxic impurities in drug substances. *Regul. Toxicol. Pharmacol. 57*, 300-306.

Bernauer, U., Heinemeyer, G., Heinrich-Hirsch, B. et al. (2008). Exposure-triggered reproductive toxicity testing under the REACH legislation: A proposal to define significant/relevant exposure. *Toxicol. Lett. 176*, 68-76.

Blaauboer, B. J. and Andersen, M. E. (2007). The need for a new toxicity testing and risk analysis paradigm to implement REACH or any other large scale testing initiative. *Arch. Toxicol. 81*, 385-387.

Boekelheide, K. and Campion, S. N. (2010). Toxicity testing in the 21st century: using the new toxicity testing paradigm to create a taxonomy of adverse effects. *Toxicol. Sci. 114*, 20-24.

Boobis, A. R., Cohen, S. M., Doerrer, N. G. et al. (2009). A data-based assessment of alternative strategies for identification of potential human cancer hazards. *Toxicol. Pathol. 37*, 714.

Bouvier d'Yvoire, M., Prieto, P., Blaauboer, B. J, et al. (2007). Physiologically-based Kinetic Modelling (PBK Modelling): meeting the 3Rs agenda. The report and recommendations of ECVAM Workshop 63. *ATLA 35*, 661-671.

Bremer, S. (2008). The need for realism in reproductive toxicity testing. *ATLA 36*, 717.

Bronaugh, R. L., Stewart, R. F., Congdon, E. R. et al. (1982). Methods for in vitro percutaneous absorption studies. I. Comparison with in vivo results. *Toxicol. Appl. Pharmacol. 62*, 474-80.

Carfi, M. A., Gennari, A., Malerba, I. et al. (2007). In vitro tests to evaluate immunotoxicity: A preliminary study. *Toxicol. 229*, 11-22.

CERI (2006). Draft validation report of TA assay using HeLa-hER-9903 to detect estrogenic activity. Available at: http://www.oecd.org/dataoecd/7/27/37504278.pdf

Chaundry, Q., Piclin, N., Cotterill, J. et al. (2010). Global QSAR models of skin sensitisers for regulatory purposes. *Chem. Central J. 4*, S5.

Chiba, M., Ishi, Y. and Sugiyama, Y. (2009). Prediction of hepatic clearance in human from in vitro data for successful drug development. *AAPS J. 11*, 262-276.

Combes, R., Balls, M., Curren, R. et al. (1999). Cell transformation assays as predictors of human carcinogenicity. The report and recommendations of ECVAM Workshop 39. *ATLA 27*, 745-767.

Combes, R., Balls, M., Illing, P. et al. (2006). Possibilities for a new approach to chemicals risk assessment. The report of a FRAME Workshop. *ATLA 34*, 621-649.

Combes, R. (2009). The "uEST" in vitro test for embryotoxicity – Validated and endorsed or not? *Toxicol. In Vitro 23*, 360-336.

Contrera, J. F., Kruhlak, N. L., Matthews, E. J. et al. (2007). Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR models. *Regul. Toxicol. Pharmacol. 49*, 172-182.

Corvi, R., Albertini, S., Hartung, T. et al. (2008). ECVAM retrospective validation of in vitro micronucleus test (MNT). *Mutagenesis 23*, 271-283.

Dartel, D. A. M., Pennings, J. L., de la Fonteyne, L. J. et al. (2010). Monitoring developmental toxicity in the embryonic stem cell test using differential gene expression of differentiation-related genes. *Toxicol. Sci. 116*, 130-139.

Duff, T., Carter, S., Feldman, G. et al. (2002). Transepithelial resistance and inulin permeability as endpoints in in vitro nephrotoxicity testing. *ATLA 30*, 53-59.

ECVAM (2009). Statement on the performance under UN GHS of three in vitro assays for skin irritation testing and the adaptation of the reference chemicals and defined accuracy values of the ECVAM skin irritation performance standards.

EFSA (2007). Opinion of the scientific panel on plant protection products and their residues on a request from the Commission related to the revision of Annexes II and III to Council Directive 91/414/EEC concerning the placing of plant protection products on the market – Toxicological and metabolism studies (Question N° EFSA-Q-2006-118)

Ennever, F. K., Noonan, T. J. and Rosenkranz, H. S. (1987). The predictivity of animal bioassays and short-term genotoxicity tests for carcinogenicity and non-carcinogenicity to humans. *Mutagenesis. 2*, 73-78.

Ennever, F. K. and Lave, L. B. (2003). Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. *Regul. Toxicol Pharmacol. 38*, 52-57.

Enoch, S. J., Cronin, M. T., Schultz, T. W. et al. (2008a). Quantitative and mechanistic read across for predicting skin sensitisation potential of alkenes acting via Michael addition. *Chem. Res. Toxicol. 21*, 513-520.

Enoch, S. J., Madden, J. C. and Cronin, M. T. (2008b). Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ. Res. 19*, 555-578.

Enoch, S. J., Roberts, D. W. and Cronin, M. T. (2010). Mechanistic category formation for the prediction of respiratory sensitisation. *Chem. Res. Toxicol. 23*, 1547-1555.

ESAC (2002). Statement on the use of scientifically-validated in vitro tests for embryotoxicity. ESAC Statement 3 June 2002. http://ecvam.jrc.it/

ESAC (2006). Statement on the scientific validating of the in vitro micronucleaus test as an alternative to the in vitro chromosome aberration assay for genotoxicity testing. ESAC Statement 17 November 2006. http://ecvam.jrc.it/

Fedorowicz, A., Lingyi, Z., Harshinder, S. et al. (2008). QSAR study of skin sensitisation using local lymph node assay data. *Int. J. Mol. Sci. 5*, 56-66.

Fjodorova, N., Vračko, M., Novič, M. et al. (2010). New public QSAR model for carcinogenicity. *Chemistry Central J. 4*, S3.

Fowler, P., Smith, K., Jong, J. et al. (in press). Reduction of misleading (false) positive results in mammalian genotoxicity assays. I. Choice of cell type. *Mutat Res*.

Freyberger, A., Wilson, V., Weimer, M. et al. (2010a). Assessment of a robust model protocol with accelerated throughput for a human recombinant full length estrogen receptor- binding assay: Protocol optimization and intralaboratory assay performance as initial steps towards validation. *Reprod. Toxicol. 30*, 50-59.

Freyberger, A., Weimer, M., Lofink, W. et al. (2010b). Short-term dynamic culture of rat testicular fragments as a model to assess effects on steroidogenesis-potential use and limitations. *Reprod. Toxicol. 30,* 36-43.

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. *ATLA 30*, 151-76.

Gerberick, G. F., Vassallo, J. D., Foertsch, L. M. et al. (2007). Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. *Toxicol. Sci. 97*, 417-427.

Goebel, C., Hewitt, N. J., Kunze, G. et al. (2009). Skin metabolism of aminophenols: Human keratinocytes as a suitable in vitro model to qualitatively predict the dermal transformation of 4-amino-2-hydroxytoluene in vivo. *Toxicol. Appl. Pharmacol. 235*, 114-123.

Gordon, J. D. and Clark, G. C. (2005). Submission of XDS's LUMI-CELL ER high-throughput system for screening estrogen-like chemicals for review by ICCVAM. http://iccvam.niehs.nih.gov/methods/endocrine/endodocs/ICCVAMSubmission28Jan05.pdf

Grindon, C., Combes, R., Cronin, M. T. et al. (2008). An integrated decision-tree testing strategy for repeat dose toxicity with respect to the requirements of the EU REACH legislation. *ATLA 36*, 93-101.

Hettwer, M., Reis-Fernandes, M. A., Iken, M. et al. (2010). Metabolic activation capacity by primary hepatocytes expand the applicability of the embryonic stem cell test as an alternative to experimental animal testing. *Reprod. Toxicol. 30*, 13-20.

Hewitt, P. (2009). Predictive safety biomarkers in non-clinical development. Presentation. http://www.insightpharmareports.com/uploadedFiles/Reports/Reports/Multiplex_Assays/Presentations/Hewitt_Phil.pdf

Horii, I. and Yamada, H. (2007). In vitro hepatotoxicity testing in the early phase of drug discovery. *AATEX 14, Spec. Issue*, 437-441.

Hosea, N. A., Collard, W.T., Cole, S. et al. (2009). Prediction of human pharmacokinetics from preclinical information: comparative accuracy of quantitative prediction approaches. *J. Clin. Pharmacol. 49*, 513-533.

Huang, S., Wiszniewskia, L., Derouette, J. P. et al. (2009). In vitro organ culture models of asthma. *Drug Discov. Today 6*, 137-144.

Hurtt, M., Cappon, G. D., Browning, A. et al. (2003). Proposal for a tiered approach to developmental toxicity testing for veterinary pharmaceutical products for food producing animals. *Food Chem. Toxicol. 41*, 611-619.

ICCVAM (1999). The murine local lymph node assay: A test method for assessing the allergic contact dermatitis potential of chemicals/compounds. National Toxicology Program, NIH Publication No. 99-4494; http://iccvam.niehs.nih.gov/

Inoue, T., Tanaka, K., Mishima, M. et al. (2007). Predictive in vitro cardiotoxicity and hepatotoxicity screening system using neonatal rat heart cells and rat hepatocytes. *AATEX 14, Spec. Issue*, 457-462.

Jäckh, C., Blatz, V., Guth, K. et al. (2010). Enzyme activities for xenobiotic metabolism in human reconstructed skin models. *ALTEX 27, Spec. Issue*, 60.

Jennings, P. (2010). The role of the Nrf2 pathway in nephrotoxicity. *ALTEX 27, Spec. Issue*, 60.

Jirova, D., Basketter, D., Liebsch, M. et al. (2010). Comparison of human human skin irritation patch test data with in vitro skin irritation assays and animal data. *Contact Dermatitis 62*, 109-116.

Jones, H. M., Parrott, N., Jorga, K. et al. (2006). A novel strategy for physiologically based predictions of human pharmacokinetics. *Clin. Pharmacokinet. 45*, 511-542.

Keller, D., Krauledat, M. and Scheel, J. (2009). Feasibility study to support a threshold of sensitization concern concept in risk assessment based on human data. *Arch. Toxicol. 83,* 1049-1060.

Kirkland, D. and Fowler, P. (2010). Further analysis of Ames-negative rodent carcinogens that are only genotoxic in mammalian cells in vitro at concentrations exceeding 1 mM, including retesting of compounds of concern. *Mutagenesis 25*, 539-553.

Kirkland, D. M., Aardema, M., Henderson, L. et al. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mut. Res. 7*, 70.

Knight, A. K., Bailey J. B. and Balcombe, J. (2005). Which drugs cause cancer? *Br. Med. J. 5*, 477-478.

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Rev. 3*, 711-715.

Kroes, R., Renwick, A. G., Feron, V. et al. (2007). Application of the threshold of toxicological concern (TTC) to the safety evaluation of cosmetic ingredients. *Food Chem. Toxicol. 45*, 2533-2562.

Kusama, M., Toshimoto, K., Maeda, K. et al. (2010). In silico classification of major clearance pathways of drugs based on physicochemical parameters. *Drug Metab. Dispos. 38*, 1362-1370.

La Sala, G., Farini, D., and De Felici, M. (2010). Estrogenic in vitro assay on mouse embryonic Leydig cells. *Int. J. Dev. Biol. 54*, 717-722.

Landsiedel, R., Fabian, E., Gamer, A. et al. (2010). The use of alternative methods for toxicity testing of nanomaterials. *ALTEX 27, Spec. Issue*, 78.

Langezaal, I. S., Hoffmann, S., Hartung, T. et al. (2002). Evaluation and prevalidation of an immunotoxicity test based on human whole-blood cytokine release. *ATLA 30*, 581-595.

Lankveld, D. P., Van Loveren, H., Baken, K. A. et al. (2009). In vitro testing for direct immunotoxicity: State of the art. *Meth. Mol. Biol. 598*, 401-423.

Lazzari, G., Tessaro, I., Crotti, G. et al. (2008). Developmental of an in vitro test battery for assessing chemical effects on bovine germ cells under the ReProTect umbrella. *Toxicol. Appl. Pharmacol. 233*, 360-370.

LeBoeuf, R. A., Kerckaert, G. A., Aardema, M. J. et al. (1996). The pH 6.7 Syrian hamster embryo cell transformation assay for assessing the carcinogenic potential of chemicals. *Mut. Res. Fund. Mol. Mech. Mutagen. 356*, 85-127.

Long, M. E. (2007). Predicting carcinogenicity in humans: The need to supplement animal-based toxicology. *AATEX 14*, *Spec. Issue*, 553-559.

Luciano, A. M., Franciosi, F., Lodde, V. et al. (2010). Transferability and inter-laboratory variability assessment of the in vitro bovine oocyte maturation (IVM) test within ReProTect. *Reprod. Toxicol. 30*, 81-88.

Malygin, V. V., Sokolov, V. B., Richardson, R. J. et al. (2003). Quantitative structure-activity relationships predict the delayed neurotoxicity potential of a series of o-alkyl-o-methylchloroformimino phenylphosphonates. *J. Toxicol. Environ. Health A 66*, 611-25.

Manga, N., Duffy, J. C., Rowe, P. H. et al. (2003). A hierarchical QSAR model for urinary excretion of drugs in humans as a predictive tool for biotransformation. *QSAR Combinat. Sci. 22*, 263-273.

Martinozzi, S. (2010). Skin sensitisation and combination of in vitro tests. Oral presentation held at 16th Congress on Alternatives to Animal Testing, Linz, Austria, 1-4 September 2010.

Maunz, A. and Helma, C. (2008). Prediction of chemical toxicity with local support vector regression and activity-specific kernels. *SAR QSAR Environ. Res. 19*, 413-431.

Mauthe, R. J., Gibson, D. P., Bunch, R. T. et al. (2001). The Syrian hamster embryo (SHE) cell transformation assay: review of the methods and results. *Toxicol. Pathol. 29*, 138.

McKim, J. M. Jr. (2010). Building a tiered approach to in vitro predictive toxicity screening: a focus on assays with in vivo relevance. *Combinat. Chem. High Throughput Screen. 13*, 188-206.

Morin, J. P., De Broe, M. E., Pfaller, W. et al. (1997). Nephrotoxicity testing in vitro: the current situation. ECVAM Nephrotoxicity Task Force Report 1. *ATLA 25*, 497-503.

Munro, I. C., Ford, R. A., Kennepohl, E, et al. (1996). Thresholds of toxicological concern based on structure-activity relationships. *Drug Metab. Rev. 28*, 209-217.

Natsch, A., Gfeller, H., Rothaupt, M. et al. (2007). Utility and limitations of a peptide reactivity assay to predict fragrance allergens in vitro. *Toxicol. In Vitro 21*, 1220-1226.

O'Brien, P. J., Irwin, W., Diaz, D. et al. (2006). High concordance of drug induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. *Arch. Toxicol. 80*, 580-604.

OECD (2004). OECD Guideline for the testing of chemicals. Skin absorption in vitro method. TG 428, Adopted 13 April 2004.

OECD (2007). Detailed review paper on cell transformation assays for detection of chemical carcinogens (Draft 4th version). OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 31.

OECD (2009). Multi-laboratory validation report of the h295r steroidogenesis assay to identify modulators of testosterone and estradiol production, No 132, Paris, France.

OECD (2010). Guidelines for the testing of chemicals, Section 4: Health effects. Test No. 417: Toxicokinetics.

Olson, H., Betton, G., Robinson, D. et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and animals. *Reg. Toxicol. Pharmacol. 32*, 56-67.

Paquette, J. A., Kumpf, S. W., Streck, R. D. et al. (2008). Assessment of the embryonic stem cell test and application and use in the pharmaceutical industry. *Birth Defects Res. B Dev. Repro. Toxicol. 83*, 104-111.

Parry, J. M., Parry, E., Phrakonkham, P. et al. (2010). Analysis of published data for top concentration considerations in mammalian cell genotoxicity testing. *Mutagenesis 25*, 531-538.

Pauwels, M., Dejaegher, B., Vander Heyden, Y. et al. (2009). Critical analysis of the SCCNFP/SCCP safety assessment of cosmetic ingredients (2000-2006). *Food Chem. Toxicol. 47*, 898-905.

Pessina, A., Albella, B., Bueren, J. et al. (2001). Prevalidation of a model for predicting acute neutropenia by colony forming unit granulocyte/macrophage (CFU-GM) assay. *Toxicol. In Vitro 15*, 729-740.

Peters, A. K., Steemans, M., Hansen, E. et al. (2008a). Evaluation of the embryotoxic potency of compounds in a newly revised high throughput embryonic stem cell test. *Toxicol. Sci. 105*, 342-350.

Peters, A. K., Wouwer, G. V., Weyn, B. et al. (2008b). Automated analysis of contractility in the embryonic stem cell test, a novel approach to assess embryotoxicity. *Toxicol. In Vitro 22*, 1948-1956.

Poulin, P. and Theil, F. P. (2002). Predictions of pharmacokinetics prior to in vivo studies: I. Mechanism based prediction of volume if distribution. *J. Pharm. Sci. 91*, 129-156.

Prieto, P., Baird, A. W., Blaauboer, B. J. et al. (2006). The assessment of repeated dose toxicity in vitro: a proposed approach. The Report and Recommendations of ECVAM Workshop 56. *ATLA 34*, 315-341.

Python, F., Goebel, C. and Aeby, P. (2007). Assessment of the U937 cell line for the detection of contact allergens. *Toxicol. Appl. Pharmacol. 220*, 113-124.

Roberts, D. W. and Aptula, A. O. (2008). Determinants of skin sensitisation potential. *J. Appl. Toxicol. 28*, 377-387.

Roberts, D. W. and Patlewicz, G. Y. (2010). Updating the skin sensitization in vitro data assessment paradigm in 2009 – a chemistry and QSAR perspective. *J. Appl. Toxicol. 30, 286-288*; discussion 289.

Rogiers, W. and Pauwels, M. (2008). Safety assessment of cosmetics in Europe. *Current Problems in Dermatology 36*. Basel, Switzerland: S. Karger.

Russell, W. and Burch, R. (1959). *The principles of humane experimental technique*. London, UK: Methuen.

Sakaguchi, H., Ashikaga, T., Miyazawa, M. et al. (2009). The relationship between CD86/CD54 expression and THP-1 cell viability in an in vitro skin sensitization test – human cell line activation test (h-CLAT). *Cell Biol. Toxicol. 25*, 109-126.

SCCNFP (1999). Final: basic criteria for the in vitro assessment of percutaneous absorption of cosmetic ingredients, adopted by the SCCNFP, 23 June 1999, SCCNFP/0167/99.

SCCNFP (2000). Opinion on the Murine Local Lymph Node Assay (LLNA) adopted by the SCCNFP during the 12th plenary meeting of 3 May 2000. http://ec.europa.eu/health/scientific_committees/consumer_safety/opinions/sccnfp_opinions_97_04/sccp_out114_en.htm

SCCP (2006). Notes of guidance for the testing of cosmetic ingredients and their safety evaluation, sixth revision. SCCP/1005/06.

Schaafsma, G., Kroese, E. D., Tielemans, E. L. et al. (2009). REACH, non-testing approaches and the urgent need for a change in mind set. *Regul. Toxicol. Pharmacol. 53*, 70-80.

Schäfer-Korting, M., Bock, U., Diembeck, W. et al. (2008). The use of reconstructed human epidermis for skin absorption testing: Results of the validation study. *ATLA 36*, 161-187.

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*, 200-218.

Schmelzer, E., Mutig, K., Schrade, P. et al. (2009). Effect of human patient plasma ex vivo treatment on gene expression and progenitor cell activation of primary human liver cells in multi-compartment 3D perfusion bioreactors for extra-corporeal liver support. *Biotechnol. Bioeng. 103*, 817-827.

Schwengberg, S., Kettenhofen, R., Bohlen, H. et al. (2004). Predictive in vitro assays for cardiotoxicity and myelotoxicity of kinase inhibitors. Poster http://www.axiogenesis.com/cms/upload/applications/Downloads/Poster/090604_poster-WPC_TKI.pdf

SEC (2004). Commission Staff Working Document. Timetables for the phasing-out of animal testing in the framework of the 7th Amendment to the Cosmetics Directive (Council Directive 76/768/EEC). Brussels, 1.10.2004, 1210.

Selderslaghs, I. W., Van Rompay, A. R., De, C. W. et al. (2009). Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo. *Reprod. Toxicol. 28*, 308-320.

Sonneveld, E., Riteco, J. A., Jansen, H. J. et al. (2006). Comparison of in vitro and in vivo screening models for androgenic and estrogenic activities. *Toxicol. Sci. 89*, 173-187.

Sonneveld, E., Pieterse, B., Schoonen, W. G. et al. (2011). Validation of in vitro screening models for progestagenic activities: Inter-assay comparison and correlation with in vivo activity in rabbits. *Toxicol. In Vitro 25*, 545-554.

Søsted, H., Basketter, D. A., Estrada, E. et al. (2004). Ranking of hair dye substances according to predicted sensitization potency: quantitative structure-activity relationships. *Contact Dermatitis 51*, 241-254.

Spanhaak, S., Cook, D., Barnes, J. et al. (2008). Species concordance for liver injury: from the safety intelligence program board (6pp.). Cambridge, UK: BioWisdom, Ltd. http://www.biowisdom.com/files/SIP_Board_Species_Concordance.pdf (accessed 24.08.09).

Speit, G. (2009). How to assess the mutagenic potential of cosmetic products without animal tests? *Mutat. Res. 678*, 108-112.

Spielmann, H., Genschow, E., Brown, N. A. et al. (2004). Validation of the rat limb bud micromass test in the international

ECVAM validation study on three in vitro embryotoxicity tests. *ATLA 32*, 245-274.

Spielmann, H. (2009). The way forward in reproductive/developmental toxicity testing. *ATLA 37*, 641-656.

Spielmann, H. (2010). Editorial: The EU Commission's Draft Report on alternative (nonanimal) methods for cosmetics testing: Current status and future prospects – 2010: A missed opportunity. *ATLA 38*, 339-343.

Sreejit, P., Kumar, S. and Verma, R. S. (2008). An improved protocol for primary culture of cardiomyocyte from neonatal mice. *In Vitro Cell. Devel. Biol. – Anim. 44*, 45-50.

Tilaoui, L., Schilter, B., Tran, L.-A. et al. (2007). Integrated computational methods for prediction of the lowest observable adverse effect level of food-borne molecules. *QSAR Combinat. Sci. 26*, 102-108.

Van der Burg B., Winter, R., Man, H. Y. et al. (2010). Optimization and prevalidation of the in vitro AR CALUX method to test androgenic and antiandrogenic activity of compounds. *Reprod. Toxicol. 30*, 18-24.

Weed, D. (2005). Weight of evidence: a review of concept and methods. *Risk Anal. 25*, 1545-1557.

West, P. R., Weir, A. M., Smith, A. M. et al. (2010). Predicting human developmental toxicity of pharmaceuticals using human embryonic stem cells and metabolomics. *Toxicol. Appl. Pharmacol. 247*, 18-27.

Westmoreland, C., Carmichael, P., Dent, M. et al. (2010). Assuring safety without animal testing: Unilever's ongoing research programme to deliver novel ways to assure consumer safety. *ALTEX 27*, 61-65.

Witters, H., Freyberger, A., Smits, K. et al. (2010). The assessment of estrogenic or anti-estrogenic activity of chemicals by the human stably transfected estrogen sensitive MELN cell line: Results of test performance and transferability. *Reprod. Toxicol. 30*, 60-72.

Yamazaki, S., Skaptason, J., Romero, D. et al. (2011). Prediction of oral pharmacokinetics of cMet kinase inhibitors in humans: physiologically based pharmacokinetic model versus traditional one compartment model. *Drug Metab. Dispos. 39*, 383-393.

Zeiger, E. (1998). Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: premises, promises, and performance. *Regul. Toxicol. Pharmacol. 28*, 85-95.

Zeilinger, K., Schreiter, T., Darnell, M. et al. (2011). Scaling down of a clinical three- dimensional perfusion multicompartment hollow fiber liver bioreactor developed for extracorporeal liver support to an analytical scale device useful for hepatic pharmacological in vitro studies. *Tissue Eng. Part C Methods*, Epub ahead of print, doi: 10.1089/ten.tec.2010.0580.

Zuang, V., Barroso, J., Bremer, S. et al. (2010). ECVAM technical report on the status of alternative methods for cosmetics testing (2008-2009). EUR 24413 EN – 2010.

Correspondence to

Katy Taylor, PhD
European Coalition to End Animal Experiments (ECEAE)
16a Crane Grove, London
N7 8NN, UK
e-mail: Katy.taylor@buav.org