

An economic approach to non-animal toxicity testing for skin sensitisation

Maria Leontaridou

Thesis committee**Promotor**

Prof. Dr E.C. van Ierland
Professor of Environmental Economics and Natural Resources
Wageningen University & Research

Co-Promotors

Dr S.G.M. Gabbert
Associate professor, Environmental Economics and Natural Resources Group
Wageningen University & Research

Dr R. Landsiedel
Vice President
Experimental Toxicology and Ecology
BASF SE, Ludwigshafen, Germany

Other members

Dr S. Hoffmann, seh consulting + services, Paderborn, Germany
Prof. Dr P.H. Feindt, Wageningen University & Research
Dr H. Tobi, Wageningen University & Research
Dr M. P. M. Meuwissen, Wageningen University & Research

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE)

An economic approach to non-animal toxicity testing for skin sensitisation

Maria Leontaridou

Thesis

submitted in the fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 19 April 2017
at 11 a.m. in the Aula.

Maria Leontaridou

An economic approach to non-animal toxicity testing for skin sensitisation, 151 pages

PhD Thesis, Wageningen University, Wageningen, the Netherlands (2017)

With references, with summary in English

ISBN: 978-94-6343-136-1

DOI: 10.18174/409799

Contents

| | |
|--|----|
| 1 Introduction | 1 |
| 1.1 Principles of chemicals' risk assessment and the need for improving efficiency of toxicity testing | 1 |
| 1.2 Addressing the development of efficient toxicity testing strategies: The case of skin sensitisation | 5 |
| 1.3 Problem definition, research objectives and research questions | 8 |
| 1.4 Methodology | 10 |
| 1.5 Novelty | 12 |
| 1.6 Outline of the thesis | 15 |
| 2 A review of concepts and tools for integrating information from non-animal testing methods: The case of skin sensitisation..... | 17 |
| 2.1 Introduction | 18 |
| 2.2 Criteria for developing non-animal defined approaches..... | 21 |
| 2.2.1 Aims and conceptual criteria for developing defined approaches suggested in the toxicological literature | 21 |
| 2.2.2 An economic perspective to developing optimal defined approaches | 23 |
| 2.3 Methodological approaches for data integration into defined approaches | 26 |
| 2.3.1 Qualitative approaches..... | 26 |
| 2.3.1.1 Descriptive weight of evidence (WoE) based approaches | 26 |
| 2.3.1.2 Mode of action (MoA) based approaches..... | 27 |
| 2.3.1.3 Adverse outcome pathway (AOP) based approaches | 27 |
| 2.3.2 Quantitative approaches..... | 28 |
| 2.3.2.1 Machine learning approaches..... | 28 |
| 2.3.2.2 Cost analysis approaches | 28 |
| 2.3.2.1 Decision theoretic approaches..... | 28 |
| 2.4 Evaluating defined approaches addressing skin sensitisation according to resource efficiency criteria | 29 |
| 2.5 Conclusions and recommendations..... | 35 |
| 3 Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian Value-of-Information analysis..... | 37 |
| 3.1 Introduction | 38 |
| 3.2 Method: A decision-theoretic approach to assessing the value of testing..... | 40 |
| 3.3 Application: Optimised testing strategies for assessing skin sensitisation hazard of cosmetic ingredients..... | 45 |
| 3.4 Results | 50 |
| 3.5 Discussion and conclusions | 57 |

| | |
|--|----|
| 4 The borderline range of prediction models for skin sensitisation potential assessment: Quantification and implications for evaluating non-animal testing methods' precision..... | 61 |
| 4.1 Introduction | 62 |
| 4.2 Materials and methods | 64 |
| 4.2.1 Testing methods..... | 64 |
| 4.2.1.1 Local lymph node assay | 65 |
| 4.2.1.2 Direct peptide reactivity assay..... | 65 |
| 4.2.1.3 ARE-Nrf2 luciferase method | 65 |
| 4.2.1.4 Human cell line activation test..... | 66 |
| 4.2.1.5 The “2 out of 3” ITS for characterising skin sensitisation potential | 66 |
| 4.2.2 Approach to quantify the borderline range (BR) | 66 |
| 4.2.3 Decision rules for identifying borderline substances tested with individual non-animal methods | 69 |
| 4.2.4 Decision rules for identifying borderline substances tested with the “2 out of 3” ITS | 71 |
| 4.3 Results | 72 |
| 4.3.1. Quantification of the borderline range (BR) for the DPRA, LuSens, the h-CLAT and the LLNA..... | 72 |
| 4.3.2 Identification of borderline substances in experimental samples tested with the non-animal testing methods DPRA, LuSens and h-CLAT, and with the animal test LLNA | 73 |
| 4.3.3 Identification of borderline substances in the experimental sample tested with the “2 out of 3” ITS | 75 |
| 4.4 Discussion | 75 |
| 4.4.1 Identification of borderline substances and implications of the BR for assessing substances' skin sensitisation potential | 75 |
| 4.4.2 Precision of non-animal testing methods compared to the LLNA..... | 76 |
| 4.4.3 Precision of the “2 out of 3” ITS | 77 |
| 4.5 Conclusions | 77 |
| 5 Uncertainties in measures of predictivity: The impact of precision, sample size and sample composition on the predictive accuracy of non-animal methods for skin sensitisation | 79 |
| 5.1 Introduction | 80 |
| 5.2 Materials and methods | 82 |
| 5.2.1 Non-animal testing methods for assessing skin sensitisation potential | 82 |
| 5.2.2 Quantification of the borderline range..... | 83 |
| 5.2.3 Calculation of testing method's accuracy metrics..... | 84 |

| | |
|--|-----|
| 5.2.4 Scenarios for analysing the impact of limited precision, sample size and sample composition on non-animal methods' predictive accuracy | 84 |
| 5.3 Results | 88 |
| 5.3.1 Impact of precision uncertainty on accuracy metrics of the DPRA, LuSens, the h-CLAT and the "2 out of 3" ITS | 88 |
| 5.3.2 Impact of uncertainty in sample composition and precision on accuracy metrics..... | 89 |
| 5.3.3 Assessing the joint impact of uncertainty in sample size, sample composition and precision on accuracy metrics..... | 90 |
| 5.4 Discussion | 93 |
| 5.4.1 Precision uncertainty | 93 |
| 5.4.2 Impact of uncertainty in sample composition and precision on non-animal methods' predictivity | 94 |
| 5.4.3 Impact of uncertainty in sample composition, sample size and precision on non-animal methods' predictivity..... | 94 |
| 5.5 Conclusions | 95 |
| 6 Synthesis..... | 99 |
| 6.1 Answers to the research questions..... | 99 |
| 6.2 General discussion..... | 107 |
| 6.2.1 Wider context of the thesis | 107 |
| 6.2.2 Methodological and modelling approaches..... | 109 |
| 6.2.3 Policy relevance..... | 112 |
| 6.3 Limitations of the thesis..... | 114 |
| 6.3.1 Scope..... | 114 |
| 6.3.2 Assumptions | 115 |
| 6.3.3 Data availability..... | 116 |
| 6.4 Conclusions | 117 |
| 6.5 Suggestions for further research | 118 |
| Appendix A..... | 119 |
| Appendix B..... | 122 |
| Appendix C..... | 130 |
| Appendix D | 137 |
| Summary | 139 |
| Acknowledgements..... | 143 |
| About the author | 144 |
| References | 145 |

Alphabetical list of abbreviations:

| | |
|-----------|--|
| ACD | Allergic Contact Dermatitis |
| ANN | Artificial Neural Network |
| AOP | Adverse Outcome Pathway |
| ARE-Nrf2 | Antioxidant Response Element - Nuclear Factor Erythroid 2 |
| BN | Bayesian Networks |
| BR | Borderline Range |
| CART | Classification and regression trees |
| CBA | Cost benefit analysis |
| CEA | Cost-effectiveness analysis |
| CEFIC | European Chemical Industry Council |
| CI | Confidence Intervals |
| CLP | Classification Labelling Packaging |
| DA | Defined approaches |
| DPRA | Direct Peptide Reactive Assay |
| EC | European Commission |
| ECHA | European Chemicals Agency |
| ECVAM | European Centre for Validation of Alternative Methods |
| EU | European Union |
| EUR-ECVAM | European Union Reference Laboratory for alternatives to animal testing |
| EVTI | Expected Value of Test Information |
| FI | Fold induction |
| FN | False negative |
| FP | False positive |
| h-CLAT | Human Cell Line Activation Test |
| IATA | Integrated Approaches to Testing and Assessment |
| ICCVAM | Interagency Coordinating Committee on the Validation of Alternative Methods |
| ITS | Integrate Testing Strategies |
| KE | Key event |
| LLNA | Local Lymph Node Assay |
| MI | Mutual Information |
| MoA | Mode of action |
| NB | Naïve Bayes algorithms |
| OECD | Organisation of Economic Co-operation and Development |
| QSAR | Quantitative Structure-Activity Relationship |
| REACH | Registration, Evaluation, Authorisation and Restriction of Chemicals |
| ROC | Receiver operating characteristic |
| SI | Stimulation index |
| STS | Sequential Testing strategies |
| SVM | Support vector machines |
| TN | True negative |
| TP | True positive |
| UN-GHS | United Nations - Globally Harmonized System of Classification and Labelling of Chemicals |
| VOI | Value of Information |
| WHO | World Health Organisation |
| WoE | Weight of Evidence |

1 Introduction

1.1 Principles of chemicals' risk assessment and the need for improving efficiency of toxicity testing

Chemicals are ubiquitous in our daily lives. The use of chemicals in several products and applications, such as food products, pharmaceuticals or cosmetics, creates great benefits to society. At the same time, humans and the environment are exposed to chemicals via a number of pathways. Depending on exposure concentrations, some chemicals can have harmful effects and can pose risks to human health and the quality of the environment (WHO, 2016). The production of chemicals is expected to continue to increase in the coming years which may also increase human and environmental exposure and, consequently, the risks of adverse effects (OECD, 2012a). Controlling the risks from chemicals' use and enforcing effective control strategies are key tasks of regulatory agencies, for example, the European Chemicals Agency (ECHA), the Organisation of Economic Co-operation and Development (OECD), and the World Health Organisation (WHO), who have put a lot of effort on the development of risk assessment and risk management processes. For instance the OECD, has been developing control strategies for consumer products such as risk assessment approaches in order to ensure harmonised strategies worldwide (OECD, 2016a). Likewise, the WHO and the Food and Agriculture Organization of the United Nations (FAO) have developed safety and control measures for pesticides (WHO/FAO, 2016a; WHO/FAO, 2016b).

Risk assessment of chemicals describes the process by which information about the hazard identification of chemicals, the dose-response relationship (effects assessment), and exposure is collected and combined in order to characterise a chemical's risks (WHO, 2004; van Leeuwen et al., 2007), see Figure 1.1. Based on the risk characterisation of chemicals, risk management measures for the control of risks to human health and the environment, can be adopted in order to allow the safe use of substances. The outcomes of risk management measures depend on information from the risk assessment of chemicals, but also consider the economic relevance of a chemical, the effectiveness of a measure, or its practicality, consistency and public acceptability (Krewski et al., 2009; Gabbert and Weikard, 2010; Tralau et al., 2015). Typical risk management measures are classification and labelling, safety standards, adjustments of the production technologies, restriction of use, or even a ban of the chemical compound (Hansen and Blainey, 2008).

To ensure the protection of human health and the environment, the European Commission (EC) has established regulatory frameworks guiding the risk assessment of chemicals produced, manufactured or imported within the EU. In June 2007 the new

Introduction

chemicals' legislation "Registration, Evaluation, Authorisation and Restriction of Chemicals" (REACH) entered into force (EC, 2006). The key aims of REACH are: (i) *"...to ensure a high level of protection of human health and the environment, including the promotion of alternative methods for assessment of hazards of substances..."* (see Article 1 in EC, 2006) (ii) *"the sharing and joint submission of information... in particular information related to the intrinsic properties of substances"* (see Article 25 in EC, 2006) (iii) *"...to ensure the good functioning of the internal market while assuring that the risks from substances of very high concern are properly controlled and that these substances are progressively replaced by suitable alternative substances ..."* (see Article 55 in EC, 2006).

To meet the information requirements defined by the REACH legislation, it is estimated that about 143,000 substances need to be tested, which is much more than the initial EU estimate of approximately 30,000 substances (Schoeters, 2010). Obviously, the traditional approach for assessing chemicals' risks and hazards based on a "check-list" approach, where information about chemicals' hazard is generated performing highly standardised *in vivo* or *in vitro* testing methods for every toxicological endpoint, is not able to fill information gaps within the deadlines defined in the REACH legislation. Furthermore, generating information about the hazardous properties of chemicals is resource consuming (Koch and Ashford, 2006; Bottini and Hartung, 2009). It has been estimated that fulfilling information requirements defined by REACH – if based on existing animal tests – would increase testing costs by several billions of Euros, depending on the toxicological endpoint and the number of chemicals that need to be tested for a given endpoint (Rovida and Hartung, 2009). The need to fill information gaps for large numbers of chemicals at low cost has stimulated research on developing new and efficient approaches to toxicity testing (Schaafsma et al., 2009; Andersen and Krewski, 2010; Hartung, 2010a).

In addition to protecting human health and the environment through generating sufficient and adequate information, the REACH legislation emphasises the need to reduce animal testing (EC, 2003b; Hartung, 2010b). Article 25 of REACH legislation states that animal tests should be used as "a last resort", supporting the "3R's principle" i.e. the refinement, reduction and replacement of animal tests (Russell, 1959). Besides the REACH legislation, the European Cosmetics Regulation (EC, 2009), which was commenced in July 2013 and replaced the Cosmetics Directive (EC, 2003a), has adopted concrete steps to phase-out animal testing. Specifically, since 2009 the Cosmetics Regulation has prohibited to test finished cosmetic products on animals. Additionally, since July 2013 a full marketing ban of cosmetic products with ingredients tested in animals has been established. This stimulated the development of

testing methods which are not only able to generate relevant and sufficient information fast and less costly which replace animal testing.

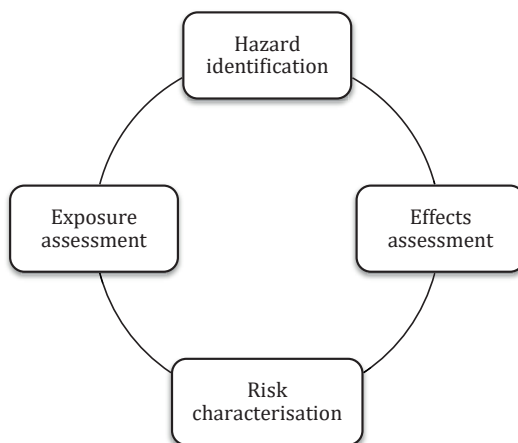


Figure 1.1: Steps of risk assessment of chemicals (adapted from van Leeuwen et al., 2007)

Furthermore, it has been increasingly acknowledged that none of the available testing methods, including the animal tests, provides perfect information about a substance's hazardous properties. Thus, information from testing is uncertain and consequently, there is a "cost of making errors". An erroneous release of a hazardous substance can cause health damages resulting in costs to the society (Hartung, 2010a). Likewise, a false classification of a safe substance may prevent the realisation of marketing benefits (Grandjean, 2015). There is a need (i) to accelerate risk assessment of chemicals in order to fill information gaps for large numbers of chemicals according to the requirements of REACH legislation, (ii) to replace animal testing, and (iii) to account for the limitations of animal tests and standalone non-animal testing methods for assessing chemicals' hazardous properties. Therefore, several concepts for integrated approaches to testing have been proposed as an innovative solution for the improvement of toxicity testing of chemicals (Grindon et al., 2006a; Nordberg et al., 2008; Jaworska and Hoffmann, 2010; Balls et al., 2012).

Initially, a testing strategy integrating information from different testing methods, was defined as a *"flexible sequence of steps ... covering the characterisation of the substance, the analysis of modes of action, the identification of possible analogues, and the evaluation of existing in vivo and in vitro testing data as well as of QSAR results"* (Ahlers et al., 2008). By sequentially combining different testing methods in a testing strategy has been considered to allow for exploiting information about the hazardous properties of a substance from various sources, thus maximising information gains, while minimising or even avoiding animal use,

testing time, and costs (Jaworska et al., 2010). Early examples of testing strategies have been developed for various toxicological endpoints, such as developmental and reproductive toxicity (Grindon et al., 2008a), eye irritation (Grindon et al., 2008b), skin corrosion (Grindon et al., 2008c), repeated dose toxicity (Grindon et al., 2008d), skin sensitisation (Grindon et al., 2008e). Whereas early studies addressing the development of testing strategies took the form of qualitative decision flow charts, there has been a scientific discussion on suitable criteria and principles for the development of testing strategies. Specific attention has been given to developing testing strategies using non-animal testing methods (Hartung et al., 2013; Rovida et al., 2015). In June 2016, the OECD published two reports (OECD, 2016b; OECD, 2016c) in which concepts related to testing strategies and risk assessment of chemicals were presented (see Table 2.1 in Chapter 2 of this thesis).

The overall debate and the process of developing non-animal testing strategies have been characterised by fundamental conceptual questions addressing the principles of data integration for assessing hazards and risks of chemicals. In particular, the fundamental questions of (i) how to integrate information from different sources in a transparent and coherent way, (ii) how to update information gains across testing sequential steps, (iii) how to quantify and reduce uncertainty across sequential steps, (iv) with which testing method to start a testing sequence, and (v) when to stop testing, have been discussed (Jaworska and Hoffmann, 2010; Jaworska et al., 2011). Answering, however, these questions, is ultimately, an economic optimisation problem (Gabbert and Weikard, 2013). Exploring the possible optimal allocation of scarce resources in order to maximise output (e.g. social welfare) is a key economic principle called “efficiency” (Hurley et al., 2000; Wynand et al., 2000). Economics as a discipline in social sciences offers a set of approaches and tools to analyse the trade-offs between competing objectives, such as information gains from testing, testing costs, and animal welfare. The aim is to increase efficiency of testing strategies and avoid unnecessary testing and costs (Nordberg et al., 2008).

Although the scientific and policy debate about the development of a “new toxicity testing paradigm” (Krewski et al., 2010), including the minimisation of animal tests, have been clearly driven by efficiency considerations, only a few studies approach the question of how to optimise toxicity testing, and the development of testing strategies, from an economic perspective. Cost-effectiveness analysis (CEA) has been applied in order to evaluate testing methods with regard to their information outcome expressed in terms of a testing method’s predictive accuracy and animals saved (Lave et al., 1988; Gabbert and van Ierland, 2010; Norlén et al., 2014). Other studies have addressed the optimisation of testing methods using

tiered approaches or sequences of testing methods from an expected utility theory perspective (Hansson and Rudén, 2007), and by using a Value of Information (VOI) analysis (Yokota et al., 2004; Yokota and Thompson, 2004; Gabbert and Weikard, 2013). However, these studies include animal tests and usually assume a pre-defined order of testing methods.

1.2 Addressing the development of efficient toxicity testing strategies: The case of skin sensitisation

During the past decade, a lot of attention has been dedicated to developing non-animal testing methods and testing strategies for the toxicological endpoint skin sensitisation (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Hartung et al., 2013; Rovida et al., 2015; Jaworska, 2016). Skin sensitisation is the toxicological endpoint assessing a substance's ability to cause allergic contact dermatitis (ACD) (UNECE, 2011). ACD is responsible for causing contact allergies affecting 15% – 20 % of the human population at least once in a lifetime (Thyssen et al., 2007). Within REACH, assessing skin sensitisation is mandatory for all chemicals produced or imported in tonnage larger than one tone per year (EC, 2006), and it is a mandatory endpoint for all substances used in cosmetic products (EC, 2009). In particular, the development of testing strategies for the assessment of skin sensitisation potential (i.e. the classification of substances as “sensitisers/non-sensitisers”), and the assessment of skin sensitisation potency (i.e. the assessment of the concentration dependent severity of an adverse effect by characterising substances as low, moderate, strong or extreme sensitisers), has been advanced from qualitative flow charts (e.g. Grindon et al. (2008e)) to deterministic approaches such as the tiered testing strategies (van der Veen, et al., 2014), or integrated testing strategies (ITS) approaches (Bauch et al., 2012; Urbisch et al., 2015a) and to probabilistic, quantitative approaches such as the Bayesian networks (Jaworska et al., 2011; Jaworska et al., 2013). Skin sensitisation has become particularly relevant for exploring and developing non-animal testing methods and for integrating information into testing strategies, among other toxicological endpoints such as endocrine disruption or liver toxicity endpoints (Valérie Zuang, 2015).

First, skin sensitisation is an economically relevant endpoint because the number of people suffering from ACD has been increasing world-wide for many years (Thyssen et al., 2007). Second, ACD causes high direct and indirect costs to society (ECHA, 2014b). Several studies have assessed direct costs arising from ACD, e.g. costs for medical treatment (Augustin and Zschocke, 2001; Ricci et al., 2006; Sætterstrøm et al., 2014), and indirect costs such as loss of well-being and productivity (Hongbo et al., 2005; Nijsten, 2012). Third, skin sensitisation testing is costly. Conservative estimates on testing costs for assessing the skin sensitisation of

Introduction

chemicals under REACH suggest 162.8 billion Euros for the total number of chemicals, considering a moderate scenario on the animal tests requirements (Rovida and Hartung, 2009). Under the same moderate scenario, animal tests for skin sensitisation, such as the Local Lymph Node Assay (LLNA) (Kimber et al., 1994; Kimber et al., 2001) described in the OECD TG 429 (OECD, 2010) and guinea pig based tests described in the OECD TG 406 (OECD, 1992), are estimated to require the use of about 823,891 animals in order to meet the requirements of REACH (Rovida and Hartung, 2009). To achieve cost minimisation in the development of testing strategies, it is important to further study the integration of information about direct testing costs, animal welfare considerations and indirect costs from the increasing occurrence of ACD, the clinically relevant effect of skin sensitisation, in the development of testing strategies for assessing skin sensitisation.

Finally, it has been acknowledged that the precision of testing methods, including animal tests, is limited. As shown for the case of skin sensitisation potential assessment, technical and biological variability can lead to misclassifications (Kolle et al., 2013; Hoffmann, 2015; Dimitrov et al., 2016; Dumont et al., 2016). The impact of biological and technical variability of the LLNA on the misclassification of substances, is associated with the fact that the classification of substances as “sensitisers/non-sensitisers” is based on clear-cut thresholds (Hoffmann, 2015). Biological and technical variability have an impact of misclassifications when test results fall within the area around the classification threshold. This area has been defined by (Kolle et al., 2013) the borderline range, also called “grey zone” by (Dimitrov et al., 2016), in which test results are discordant (also described as ambiguous, inconclusive, or borderline). Biological and technical variability can limit the precision of non-animal testing methods (Leontaridou et al., 2017a). Finally, limited precision due to the biological and technical variability can add to uncertainties underlying to measures of non-animals testing methods’ predictive accuracy (Worth and Cronin, 2001a). In particular, uncertainties occur due to variations of the size and composition of substances’ samples used to assess predictive accuracy. However, research to unravel the joint effect of different types of uncertainty on predictive accuracy metrics of non-animal testing methods is still lacking.

Given both the economic and the toxicological relevance of the skin sensitisation endpoint, several non-animal methods for assessing skin sensitisation potential have been developed in recent years (Mehling et al., 2012; Reisinger et al., 2015). The Direct Peptide Reactive Assay (DPRA) (Gerberick et al., 2004; Gerberick et al., 2007), the Antioxidant Response Element - Nuclear Factor Erythroid 2 (ARE-Nrf2) luciferase testing methods covered by KeratinoSens™ (Emter et al., 2010; Natsch et al., 2011) and the Human Cell

Activation Test (h-CLAT) (Ashikaga et al., 2006; Sakaguchi et al., 2006; Ashikaga et al., 2010; Sakaguchi et al., 2010), which have been validated by the European Centre for Validation of Alternative Methods (ECVAM; Italy). The DPRA is described in the testing guideline (TG) OECD TG 442C (OECD, 2015a), the ARE-Nrf2 luciferase method in the OECD TG 442D (OECD, 2015b) and the h-CLAT in the OECD TG442E (OECD, 2016d). The ARE-Nrf2 method is also covered by LuSens (Ramirez et al., 2014; Ramirez et al., 2016), which is currently under validation by ECVAM. In addition, we include the *in-silico* method called the OECD toolbox QSAR method which is developed by the OECD (OECD, 2012d) following the guidance document on the principles of Quantitative Structure-Activity Relationship (QSAR) models validation (OECD, 2007), in our analysis.

Individual non-animal testing methods cover different “key events” of the adverse outcome pathway (AOP) (OECD, 2012b; OECD, 2012c) of skin sensitisation, which describes the sequence of biological events and their linkages leading to the expression of an adverse outcome (i.e. an ACD incident). Since non-animal testing methods are not considered suitable to provide sufficient information to draw conclusions upon the skin sensitisation potential of chemicals (Mehling et al., 2012; Reisinger et al., 2015), a solution suggested is to integrate information from different sources and testing methods by using hypothesis-based approaches such as Bayesian networks (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Hartung et al., 2013; Jaworska et al., 2013; Jaworska, 2016) or deterministic ITS approaches (Bauch et al., 2012; Urbisch et al., 2015a). For the development of integrated strategies assessing skin sensitisation potential and potency (Jaworska, 2016) different sets of criteria have been proposed including transparency, coherency, ambiguity, or cost effectiveness (Hartung et al., 2013; Rovida et al., 2015) which are applicable also for other toxicological endpoints (Jaworska et al. 2010). It has also been suggested that the combination of individual non-animal testing methods into batteries or sequential strategies should be guided by the skin sensitisation AOP (Vinken, 2013; Patlewicz et al., 2014). The OECD has suggested the use of AOP as a guiding tool for the development of the “Integrated Approaches to Testing and Assessment” (IATA) (OECD, 2008). As a guiding tool, the AOP aims at integrating information from different testing methods for testing strategies and overall assessment of substances. In 2015, the European Chemical Industry Council (CEFIC) made a selection of IATA cases assessing skin sensitisation as validation reference (CEFIC, 2015). Besides these efforts to develop strategies for the assessment of skin sensitisation potential and potency (see Chapter 2 of this thesis for a review), the concept of “Defined Approaches” (DA), as individual

information sources to be used in IATA for skin sensitisation, was recently suggested (OECD, 2016c).

For skin sensitisation potential and potency assessment criteria for developing non-animal testing strategies have been suggested in the literature (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Hartung et al., 2013; Rovida et al., 2015; Jaworska, 2016). Moreover, the need for efficient testing, i.e. balancing information gains against costs (Gabbert and Weikard 2013; Norlén et al., 2014; Hartung et al., 2013; Rovida et al., 2015), has been repeatedly emphasised. Still, an approach to optimising non-animal testing strategies and a comprehensive evaluation of their efficiency in relation to animal tests have not become available so far. Moreover, systematic assessments of non-animal testing methods' precision and the impact of precision constraints on the predictive accuracy of these testing methods need to be further elaborated.

1.3 Problem definition, research objectives and research questions

In light of the challenges and knowledge gaps discussed above, this thesis addresses the problem how non-animal testing strategies can be optimised from an economic perspective. The focus is on non-animal testing strategies for assessing skin sensitisation potential. In order to address this problem we first need to gain insights into the criteria and the current status of non-animal testing strategies for skin sensitisation proposed in the literature. Based on the insights gained from this, an approach to optimising non-animal testing strategies, which allows balancing information gains and expected losses from testing, needs to be developed and its applicability needs to be tested. In addition, the precision of information derived from non-animal testing methods for assessing skin sensitisation potential needs to be assessed, and the uncertainties in the measures of predictive accuracy of non-animal testing methods assessing skin sensitisation potential due to limited precision, variation of sample size and sample composition needs to be evaluated.

Given the problem definition explained above, two objectives can be spelled-out. This thesis aims at developing and applying an approach to the development of optimised non-animal toxicity testing strategies assessing skin sensitisation potential that explicitly allow balancing information gains and expected costs. In this thesis, costs capture both monetary expenses for conducting non-animal toxicity testing methods, and societal costs arising from the erroneous release of a substance causing skin sensitisation called "costs of making errors". In addition, it focuses on the analysis of the impact of biological and technical variability on the precision of testing methods and examines the uncertainties in measures of predictive

accuracy of non-animal testing methods due to limited precision, variation of sample size and composition. To this end, the thesis addresses the following research questions (RQ):

RQ1: What are relevant criteria guiding the development of non-animal testing strategies for skin sensitisation potential and potency assessment?

The first research question focuses on the current practices on integrating information from non-animal testing methods for the assessment of skin sensitisation potential and potency. We introduce the conceptual criteria and informational requirements, from an economic perspective, for the development of resources-efficient testing strategies. We evaluate if and how existing testing strategies for skin sensitisation assessment meet the suggested criteria.

RQ2: How can non-animal toxicity testing strategies for assessing skin sensitisation potential be optimised?

The second research question deals with the development of the conceptual framework for optimising non-animal testing strategies, for skin sensitisation potential assessment. Furthermore, the applicability of the framework needs to be tested.

RQ3: How do technical and biological variability of non-animal testing methods influence the precision of non-animal testing methods for assessing skin sensitisation potential?

The third research question focuses on the precision of information derived from non-animal testing methods assessing skin sensitisation potential. In particular, it addresses the impact of biological and technical variability on information derived from non-animal testing methods after dichotomising continuous experimental data into binary test results.

RQ4: How do limited precision, sample size and sample composition impact the predictive accuracy of non-animal testing methods for skin sensitisation?

The predictive accuracy depends on the sample size (i.e. the number of substances which were tested to determine it), the sample composition, and the precision of a testing method. Usually, non-animal methods use prediction models to transform continuous read-outs of the test into dichotomous results by applying threshold values above and below which the test substance is assessed as positive or negative. Due to intra-test variability the precision of any testing method is limited. The fourth research question focuses on exploring the impact of limited precision and uncertainties due to varying sample size and sample composition on the predictive accuracy of non-animal testing methods. The impacts are analysed individually and in combination.

The research questions are addressed in Chapters 2 to 5 of the thesis.

1.4 Methodology

In Chapter 2 we address *RQ1* by surveying the scientific literature regarding the development of non-animal toxicity testing strategies for assessing skin sensitisation. In particular, we identify key criteria suggested in the toxicological literature for the development of testing strategies aiming at combining individual non-animal testing methods. We propose the conceptual and informational criteria required to improve resource-efficiency in the development of testing strategies from an economic perspective. Furthermore, we discuss how the criteria proposed in the toxicological literature, for example coherency, transparency and ambiguity (Hartung et al., 2013; Rovida et al., 2015) have been practically implemented. We compare these criteria with the conceptual and informational criteria from an economic perspective in order to evaluate whether existing non-animal toxicity testing strategies for skin sensitisation allow for balancing information gains and losses. Based on that, we draw conclusions on whether testing strategies can be characterised as resource-efficient and we provide suggestions in order to further improving efficiency. Finally, we discuss implications raised from the evaluation of existing approaches to develop non-animal toxicity testing strategies.

In Chapter 3, we address *RQ2* by developing a decision-theoretic model for the optimisation of non-animal toxicity testing strategies for assessing skin sensitisation potential using Bayesian Value-of-Information (VOI) analysis (Hirshleifer, 1971; Olson, 1990; Howson and Urbach, 1991; Claxton 1999). Performing a testing method is assumed to have value if and only if expected social net gains from an optimal decision with additional information derived from testing outweigh expected net gains from decision-making without evidence. The expected value of test information (EVTI) can be expressed as the probability weighted sum of social net gains under posterior beliefs that the substance is hazard and non-hazard, respectively. The probability of seeing a positive or negative test result is used as weights (Yokota et al., 2004; Yokota and Thompson, 2004; Gabbert and Weikard, 2013; Leontaridou et al., 2016). Clearly, a testing method, or a testing strategy consisting of a battery of sequential combinations of non-animal testing methods, should be performed if the EVTI is positive and exceeds testing costs. Quantifying the EVTI, allows for ranking testing methods and their combinations into testing strategies, from a social welfare perspective. Bayesian VOI analysis offers a guiding tool for the construction of sequential testing strategies which allows determining the initial testing method, the ordering of testing methods required for the collection of sufficient information and when testing should stop.

The model developed in Chapter 3 is applied to a set of validated or pre-validated non-animal methods (i.e. the DPRA, the OECD Toolbox QSAR, ARE-Nrf2 luciferase method covered by KeratinoSens™ and LuSens, and the h-CLAT), seven battery combinations of these methods, and 236 sequential 2-test and 3-test strategies composed of these methods. Their EVTI net of testing costs is compared to that of the animal test LLNA and AOP based testing strategies suggested in the literature. Determining the societal gains and losses from releasing a hazardous or a non-hazardous substance requires estimating expected marketing benefits, and balancing them with expected direct and indirect health damage costs caused by ACD. The latter is assumed to be a function of the skin sensitisation prevalence of a substance (Schnuch et al., 2011; Peiser et al., 2012; Schnuch et al., 2012) within the EU. As a proof-of-concept case, the Bayesian VOI model is applied to the preservative Methylisothiazolinone which is used for the formation of Kathon CG, known which is an ingredient for cosmetic products with high skin sensitisation prevalence (Uter et al., 2013).

In Chapter 4 we address *RQ3* by assessing the influence of technical and biological variability on the precision of non-animal testing methods for assessing skin sensitisation. We analyse how the classification of a substance as “sensitiser/non-sensitiser”, by dichotomising experimental data from non-animal testing methods using clear-cut classification thresholds, can be influenced by the biological and technical variability of the testing method. Specifically, we develop a method for quantifying the range around the classification threshold of non-animal testing methods within which a method is likely to deliver discordant test results for the binary classification of substances. This range has been called “borderline range” (Kolle et al., 2013) or “grey zone” (Dimitrov et al., 2016). Acknowledging that for any testing method, including the “first choice” animal tests (i.e. LLNA), the borderline range defines the area in which binary test results, as derived by dichotomising continuous experimental data, are discordant (inconclusive, ambiguous or borderline). This offers a new perspective to the evaluation of the precision of the prediction models using classification thresholds, of testing methods: Substances for which test results fall into the borderline range can neither be classified as positive, thus indicating an adverse effect, or negative. The classification is, therefore, inconclusive which indicates the need for further information in order to draw conclusion upon the skin sensitisation potential of the tested substance.

The borderline range is determined by calculating the pooled standard deviation of experimental test results revealed from repeated testing of substances. The standard deviation was pooled across substances and concentrations. Then, the borderline range is applied as an additional classification rule together with the threshold criteria of testing

methods. In that way, we identify substances as positive (i.e. sensitisers), negative (i.e. non-sensitisers) or discordant (i.e. substances yielding test results within the borderline range). Furthermore, the percentage of substances yielding discordant test results, thus within the borderline range, is used as a measure of a non-animal testing method's precision constraint. This analysis was performed for selected non-animal testing methods used for assessing skin sensitisation potential, i.e. the DPRA, LuSens, the h-CLAT, and the "2 out of 3" Integrated Testing Strategy (ITS) formerly called "2 out of 3" weight of evidence (WoE) approach (Bauch et al., 2012; Urbisch et al., 2015a). We quantify the borderline range for sets of experimental results compared to the LLNA as the reference animal test.

In Chapter 5, we address *RQ4* by exploring the uncertainties in predictive accuracy metrics of testing methods by quantifying the impact of limited precision and variations in sample size and sample composition, on the predictive accuracy of non-animal testing methods assessing skin sensitisation potential. We analyse these impacts for selected accuracy metrics usually used to characterise the predictive accuracy of these methods. i.e. sensitivity, specificity and concordance (also called accuracy) (Krzanowski and Hand, 2009) by means of contingency tables which is a well-established way to assess the predictive accuracy of tests (Cooper et al., 1979). The impact of limited precision is determined by quantifying sensitivity, specificity and concordance based on experimental samples including substances yielding tests results within the borderline range (i.e. borderline substances), and comparing these with accuracy metrics derived from samples after borderline substances are excluded. Furthermore, the impact of variations in sample size and sample composition on predictive accuracy metrics is assessed for randomised samples of substances using the non-parametric bootstrap resampling analysis (Jones et al., 2000; Wehrens et al., 2000). We calculate the confidence limits and standard deviations of the accuracy metrics, in order to provide estimates for the uncertainty related to accuracy metrics due to the use of samples with pre-defined compositions of substances. Besides sample composition, accuracy metrics are influenced by the number of substances (sample size) used for the assessment of accuracy metrics. The impact of limited precision and variations in sample size and composition (individually and in combination) on the predictive accuracy of the non-animal testing methods is assessed using the DPRA, LuSens, the h-CLAT and the "2 out of 3" ITS.

1.5 Novelty

It has been acknowledged that there is a need to fill in the information gaps considering the hazardous properties of thousands of substances (Ahlers et al., 2008; Krewski et al., 2009; Schaafsma et al., 2009; Krewski et al., 2010). At the same time there has been increasing

support for a systematic reduction and replacement of animal testing (Kinsner-Ovaskainen et al., 2009; Daston et al., 2015). The need for adequate and sufficient information derived from fast and less costly non-animal testing methods has emerged the development of testing strategies, integrating information from individual non-animal testing methods (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Hartung et al., 2013; Rovida et al., 2015; Jaworska, 2016). From an economic perspective this requires to balance the informational gains from performing testing methods with costs such as testing costs and the costs of making errors. Obviously, balancing gains and losses from conducting toxicity testing can be considered as an economic problem.

So far, however, only few studies have become available addressing the problem how to optimise toxicity testing from an interdisciplinary perspective, linking toxicology with economics (Olson, 1990). Chapter 2 reviews existing non-animal testing strategies for the assessment of skin sensitisation potential and potency. This chapter also identifies the criteria suggested in the toxicological literature for developing non-animal testing strategies (Jaworska, 2016) and further evaluates them from the economic perspective. Chapter 2 introduces the standard economic approach by suggesting the conceptual and informational criteria necessary to establish resource-efficiency in the development of testing strategies. Although criteria for developing resource-efficient testing strategies for the assessment of skin sensitisation refer to cost effectiveness (Hartung et al., 2013; Rovida et al., 2015), this chapter offers novel insights on the practical implementation of optimisation methods for the development of resource-efficient testing strategies from an economic perspective.

Further, Chapter 3 develops a decision-theoretic framework for the optimisation of sequential testing strategies using the Bayesian VOI analysis. While the theoretical foundations of the framework, i.e. expected-utility theory and Bayesian inference, are not new per se, there exist only few applications to the problem of sequential testing, focusing exclusively on selected endpoints, i.e. carcinogenicity (Yokota and Thompson, 2004) and mutagenicity (Gabbert and Weikard, 2013), where animal tests are usually included in the testing strategies suggests. In Chapter 3, we apply for the first time the Bayesian VOI analysis to the optimisation of non-animal testing strategies for the assessment of skin sensitisation potential. We provide an evaluation of the EVTI of non-animal testing methods and strategies in comparison to the EVTI of the animal test. The Bayesian VOI model suggested in Chapter 3 guides the construction of sequential testing strategies determining which non-animal testing method should be performed first, how many non-animal testing methods should be include into the strategy and when testing should stop. Further, we examine the role of the adverse

outcome pathway of skin sensitisation as a guide for developing optimal sequential non-animal testing strategies.

Besides developing and applying the Bayesian VOI model, this thesis provides novel insights on the informational outcomes derived from conducting non-animal testing methods. In particular, we introduce a methodology for estimating the influence of technical and biological variability on the precision of non-animal testing methods, in Chapter 4. The uncertainty of information derived from toxicity testing has been frequently discussed in the toxicology literature (Paparella et al., 2013; Heringa et al., 2015). For example, the impact of technical and biological variability on test results derived from testing methods (Kolle et al., 2013; Hoffmann, 2015; Dimitrov et al., 2016; Dumont et al., 2016; Leontaridou et al., 2017a), and the effect of overfitting experimental data from testing (Kopp-Schneider et al., 2013) have been discussed. However, the estimation of non-animal testing methods' borderline range and the identification of substances yielding borderline test results provide a novel approach to classify substances, to identify discordant test results and to decide if additional information is needed.

Chapter 4 highlights the problem of using clear-cut thresholds for the binary classification of substances without considering the borderline range of testing methods. This, however, has not been sufficiently linked to the implications that may occur, if these substances (i.e. substances classified based only on clear-cut thresholds) are used to assess the predictive accuracy of testing methods. As a consequence, the impact of substances yielding test results within the borderline range of testing methods is highly relevant for drawing conclusions about a non-animal testing method's predictive accuracy, which is addressed in Chapter 5. The uncertainty related to the assessment of predictive accuracy of testing methods (Worth and Cronin, 2001a) has been previously discussed. The predictive accuracy metrics are usually calculated based on available samples of substances and they are used as point estimates, thus without accounting for uncertainties. Besides predictive accuracy being influenced by the limited precision of the testing methods, expressed in number of substances which yield test results within the borderline range; the sample size and sample composition can impact the uncertainty in accuracy metrics of non-animal testing methods. Chapter 5 examines the uncertainties underlying to the predictive accuracy metrics due to limited precision of non-animal testing methods. Furthermore, it analyses the impact of the variations in sample size and the sample composition on predictive accuracy metrics of non-animal testing methods. Finally the effect of varying sample size and sample composition in combination with the limited precision of non-animal testing methods is examined.

1.6 Outline of the thesis

The remaining thesis chapters are structured as follows (Table 1.1): Chapter 2 surveys the state-of-the-art and the construction criteria currently suggested in the toxicological literature for developing toxicity testing strategies for assessing skin sensitisation. Furthermore, suggestions for the conceptual and informational criteria for increasing the – resource-efficiency in the development of non-animal toxicity testing strategies are proposed. Chapter 3 describes the Bayesian VOI model for the evaluation of toxicity testing methods. The model is applied on the evaluation of sequential testing strategies for the assessment of skin sensitisation potential. Chapter 4 describes the quantification of the borderline range of testing methods. The borderline range is applied as an additional classification rule, next to the threshold criteria of non-animal testing methods assessing skin sensitisation potential, in order to identify the substances as positive, negative or discordant in case test results fall into the borderline range. Chapter 5 addresses the uncertainties underlying to the predictive accuracy metrics for non-animal testing methods assessing skin sensitisation due to limited precision by calculating accuracy metrics with and without considering the borderline range in the prediction models of non-animal testing methods. Further the impact of variations in sample size and sample composition on uncertainties in the predictive accuracy metrics is assessed. Chapter 6 summarises the results of this thesis answering the research questions, discusses the main findings and addresses the policy relevance of this thesis. Finally, limitations of the methods applied in this thesis are discussed, and suggestions for further research are proposed.

Table 1.1: Outline of the remaining chapters of this thesis

| Focus on the improvement of the efficiency of non-animal toxicity testing strategies | |
|---|---|
| Ch. 2 | A review of concepts and tools for integrating information from non-animal testing methods: The case of skin sensitisation. |
| Ch. 3 | Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian Value-of-Information analysis. |
| Focus on the information from non-animal testing methods and strategies | |
| Ch. 4 | The borderline range of prediction models of testing methods for skin sensitisation potential assessment: Quantification and implications for evaluating non-animal testing methods' precision. |
| Ch. 5 | Uncertainties in measures of predictivity: The impact of precision, sample size and sample composition on the predictive accuracy of non-animal methods for skin sensitisation. |
| Closing chapter | |
| Ch. 6 | Synthesis: Main findings of this thesis, general discussion, limitations of this thesis, conclusions and suggestions for further research. |

2 A review of concepts and tools for integrating information from non-animal testing methods: The case of skin sensitisation¹

Integrating information from *in vitro*, *in silico* and *in chemico* methods into non-animal toxicity testing strategies has been widely considered an innovative way of phasing-out animal testing. At the same time, non-animal testing strategies are considered to provide adequate and relevant information about chemicals' hazardous properties in a resource-efficient way. The aim of this chapter is to identify the conceptual criteria for developing resource-efficient testing strategies, and at evaluating existing testing strategies under these criteria. This chapter provides an overview of the definitions proposed in the scientific toxicological literature to characterise the process of integrating information into strategies for hazard and risk assessment of chemicals. We, further, present the general conceptual criteria which have been suggested in the scientific toxicological literature for guiding the process of data integration. Next, we propose a set of conceptual and informational criteria for combining information from different methods into strategies in a resource-efficient way. This explicitly acknowledges that resource efficiency, being is fundamental economic concept, requires balancing the gains and losses from using scarce resources. Finally, we evaluate whether existing testing strategies, addressing skin sensitisation, meet the suggested resource efficiency criteria. We conclude that existing testing strategies focus predominantly on maximising toxicity information, whereas direct and indirect testing costs (including also welfare losses for society in case of unintended health or environmental damage) are either ignored or only addressed in a non-quantitative way.

¹ Chapter 2 is based on the manuscript in preparation: Leontaridou M., Gabbert S., Landsiedel R., (2017). A review of concepts and tools for integrating information from non-animal testing methods: The case of skin sensitisation.

2.1 Introduction

Non-animal testing methods and toxicity testing strategies aim at a subsequent reduction of the number of animal tests used and, ultimately, a full replacement of animal testing (Ahlers et al., 2008; Hartung, 2010a; Adler et al., 2011; Reisinger et al., 2015). During the past decade, non-animal testing strategies have been developed for several toxicological endpoints. In addition to developing individual methods such as *in vitro* methods, *in silico* and *in chemico* methods, experts in science, industry and regulatory agencies have put a lot of effort in combining non-animal testing methods into toxicity testing strategies (Tollefsen et al., 2014). Integrating information from various sources for human health endpoints (ECHA, 2014a; Gocht et al., 2015; Worth and Patlewicz, 2016) such as skin sensitisation and eye irritation (Sauer et al., 2016), repeated dose toxicity and toxicity to reproduction (Gocht et al., 2015) as well as environmental endpoints such as fish toxicity (Nendza et al., 2014) have received a lot of attention. Basically, a testing strategy is an integrated combination of different testing methods for achieving an adequate assessment of the hazardous properties of substances. Different conceptual approaches for combining information from *in silico*, *in vitro* or *in chemico* methods have been proposed in the scientific toxicological literature. The suggested approaches have been characterised as “*Integrated Testing Strategies*” (ITS), “*Sequential Testing Strategies*” (STS), “*Weight of Evidence*” (WoE) and “*Integrated Approach to Testing and Assessment*” (IATA). Definitions of these approaches showed terminological overlaps and made a clear delineation difficult. For example, IATAs have been described as overarching “...data integration approaches... e.g. ITS, STS, WoE or other IATA strategies...” (Tollefsen et al., 2014). Furthermore, the WoE approach has been closely related to the concept of ITS or DA, with the difference being that “...the WoE approach is usually based on existing data while ITS should prospectively address which assays need to be performed ...” (Rovida et al., 2015). Hartung et al., (2013) emphasized the need to “... understand that WoE and ITS are two different concepts although they combine the same types of information! In WoE there is no formal integration, usually no strategy, and often no testing...” (Hartung et al., 2013), while the concept of ITS “...enables an integrated and systematic approach to guide testing such that the sequence...is tailored to the chemical-specific situation...adapted and optimized for meeting specific information target” (Jaworska and Hoffmann, 2010). Despite terminological differences between the WoE concept and the ITS concept, both attempt to integrate information (Balls et al., 2006; Rovida et al., 2015). In June 2016 the OECD published guidance documents to clarify terminology (OECD, 2016b; OECD, 2016c), see Table 2.1.

Table 2.1: Definition of terminologies used to describe different schemes of integrating information from different sources into strategies for hazard and risk assessment

| Schemes | Definition | Reference |
|-------------|--|-------------------------------------|
| IATA | “An Integrated Approach to Testing and Assessment is an approach based on multiple information sources used for the hazard identification, hazard characterisation and/or safety assessment of chemicals. An IATA integrates and weights all relevant existing evidence and guides the targeted generation of new data, where required, to inform regulatory decision-making regarding potential hazard and/or risk.” | Paragraph 1 OECD, (2016c) |
| WoE | “A Weight of Evidence determination means that expert judgement is applied on an ad hoc basis to the available and scientifically justified information bearing on the determination of hazard or risk” | Paragraph 4 OECD, (2016c) |
| DA | “A Defined Approach to testing and assessment consists of a fixed data interpretation procedure used to interpret data generated with a defined set of information sources, that can either be used on its own, or together with other information sources within an IATA” “Defined Approaches to testing and assessment can be designed in different ways, and may take for example the form of a Sequential Testing Strategy (STS) or an Integrated Testing Strategy (ITS)” | Paragraph 5 and 10 OECD, (2016c) |
| ITS | “An Integrated Testing Strategy is an approach in which multiple sources of data or information are assessed at the same time by applying a variety of specific methodologies to convert inputs from the different information sources into a prediction” | Paragraph 11 OECD, (2016c) |
| STS | “A Sequential Testing Strategy is a fixed stepwise approach for obtaining and assessing test data, involving interim decision steps, which, depending on the test results obtained, can be used on their own to make a prediction or to decide on the need to progress to subsequent steps. At each step, information from a single source/method is typically used by applying a prediction model associated with that source/method” | Paragraph 12 OECD, (2016c) |

This chapter focuses on DAs, which were introduced by the OECD as a component of an IATA which comprises different approaches to integrate information from computational, *in vitro* and *in chemico* testing methods, and which can be designed as ITS or STS (OECD, 2016b; OECD, 2016c).

Besides replacing or reducing animal testing, a key driver for the development of DAs has been the need to acquire sufficient and relevant information about chemicals' hazardous properties with less time and at lower costs than the traditional animal tests (Gabbert and Weikard, 2010; Gabbert and Weikard, 2013; Hartung et al., 2013; Rovida et al., 2015). Clearly, maximising the information outcomes from testing, decreasing the time needed to attain hazard and risk information, reducing costs of testing, and minimising or even avoiding the use of animals are competing objectives. Balancing competing objectives is a fundamental economic principle which guides the optimal, i.e. resource efficient, use of scarce resources in order to either maximise information outcomes at a given resource endowment, or to minimise costs for achieving a given outcome target (Clemen and Reilly, 2001). Criteria and conceptual requirements for developing resource efficient testing strategies have been

proposed in the toxicological literature (Jaworska and Hoffmann, 2010; Krewski et al., 2010; Hartung et al., 2013; Rovida et al., 2015). So far, however, a systematic review of criteria that define “resource efficient” testing, and an evaluation of DAs with regard to these criteria has not been provided.

The aims of this chapter are, therefore, twofold. First we provide a systematic review of criteria that were proposed in the scientific toxicological literature for constructing DAs to be used for hazard and risk assessment of chemicals. Second, we suggest the conceptual and informational criteria that guide resource efficient data integration into DAs. Then we evaluate existing DAs with respect to the conceptual and informational criteria. This evaluation focuses on DAs for assessing skin sensitisation potential (i.e. hazard identification) and potency (i.e. sub-categorisation into weak, moderate, strong and extreme sensitisers). Skin sensitisation is the clinically relevant endpoint for assessing allergic contact dermatitis (ACD) (UNECE, 2011). Approximately 15 - 20% of the human population suffer from an ACD incident once in their life (Thyssen et al., 2007). Assessing chemicals’ ability to cause ACD – i.e. their skin sensitisation potential or potency – is, therefore, a key requirement for the safety assessment chemicals falling under the European chemicals’ legislation REACH (EC, 2006) and the European Cosmetics Regulation (EC, 2009). Skin sensitisation can, therefore, be used as an illustrative case for the fundamental challenge on how toxicity testing should be conducted resource-efficiently and without animal use.

Chapter 2 is structured as follows. A systematic overview of the criteria suggested for developing DAs is presented in Section 2.2. Section 2.2.1 presents conceptual criteria for developing DAs which were proposed in the toxicological literature of the past decade. Section 2.2.2 suggests conceptual and informational criteria for developing optimal, i.e. resource-efficient DAs, from an economic perspective. Section 3, then, describes qualitative (Section 3.1) and quantitative (Section 2.3.2) methodological approaches used for integrating different types of information into DAs. Section 2.4 offers a detailed evaluation of DAs for skin sensitisation potential and potency assessment with regard to the conceptual and informational criteria which are suggested in Section 2.2.2. Section 2.5 concludes and provides suggestions for further improvement of the DAs to ensure resource efficient and animal-free testing.

2.2 Criteria for developing non-animal defined approaches

2.2.1 Aims and conceptual criteria for developing defined approaches suggested in the toxicological literature

The overall goal of integrating information from different sources, in particular non-animal experimental or computational methods, is to generate adequate information at low cost, and with a minimum or no animal use (Jaworska et al., 2010; Jaworska and Hoffmann, 2010; Hartung et al., 2013; Rovida et al., 2015). The integration of information is characterised as a dynamic process which progresses along with development of non-animal testing methods and the mechanistic understanding of endpoints (Tollefsen et al., 2014; Jaworska, 2016). In the scientific toxicological literature of the past decade a number of criteria have been suggested, seeking to explain “*what the ITSs should be*” (Jaworska and Hoffmann, 2010), “*what the ITS should contain*” (Rovida et al., 2015), “*what the DAs should be associated with*” (OECD, 2016b). These criteria are summarised in Table 2.2.

Criteria, thus far, focus on (i) selecting and integrating reliable and accurate information from scientifically robust sources and testing methods (ii) guiding the final interpretation of the informational outcomes from a DA into decisions about the hazardous properties of substances (iii) evaluating the performance of the selected individual sources of information and the strategies themselves. Publications discussing criteria for developing optimal testing strategies often refer to ITS, however, sets of criteria have also been proposed and discussed for IATA, see e.g. (Tollefsen et al., 2014), and for DA, see e.g. (OECD, 2016b). As explained in the introduction, we will use in the following the term DA (i.e. defined approaches) as an overarching concept.

Table 2.2: Criteria for guiding the development of defined approaches (DAs) proposed in the scientific toxicological studies of the past decade

| Criteria to be considered in the development of DAs | References |
|---|---|
| Criteria related to generating information | |
| Optimal extraction and use of information from existing data. | Schaafsma et al., 2009; Jaworska et al., 2010; Jaworska and Hoffmann, 2010; Schoeters, 2010; De Wever et al., 2012 |
| Combination of different types of information (e.g. <i>in vitro</i> , <i>in silico</i> , <i>in chemico</i> ; use of <i>in vivo</i> only if necessary). | Hoffmann et al., 2008; Dellarco et al., 2010; De Wever et al., 2012 |
| Use of scientifically valid (accurate, reliable) and adequate information. | Hoffmann and Hartung, 2006; Ahlers et al., 2008; Schaafsma et al., 2009; Krewski et al., 2010; Schoeters, 2010; De Wever et al., 2012 |
| Evaluation of the performance (e.g. predictivity, goodness-of-fit, robustness) of individual testing methods and strategies. | Jaworska and Hoffmann, 2010; Tollefsen et al., 2014; Rovida et al., 2015 |
| Criteria related to costs and resource use | |
| Reduction/minimisation of direct testing costs for generating information and indirect costs i.e. testing time, costs of misclassifications, costs for regulatory validation of testing methods and strategies. | Lewis et al., 2007; Kinsner-Ovaskainen et al., 2009; Schaafsma et al., 2009; Krewski et al., 2010; De Wever et al., 2012 |
| Protect animal welfare; reduce/minimise the number of animal used in testing or animal suffering when animal testing is considered unavoidable. | Hoffmann and Hartung, 2006; Lewis et al., 2007; Hoffmann et al., 2008; Schaafsma et al., 2009; Krewski et al., 2010 |
| Evaluation of cost-effectiveness or efficiency. | Hoffmann and Hartung, 2006; Lewis et al., 2007; Hoffmann et al., 2008; Schaafsma et al., 2009; Krewski et al., 2010 |
| Conceptual requirements for DA development | |
| Data integration and final conclusion based on a coherent methodology (e.g. unambiguous algorithm). | Jaworska and Hoffmann, 2010; De Wever et al., 2012; Basketter et al., 2013; Rovida et al., 2015 |
| Specifications on the applicability domain, endpoint assessed and regulatory purpose (hazard or potency assessment). | Kinsner-Ovaskainen et al., 2009; Krewski et al., 2010; De Wever et al., 2012; Basketter et al., 2013; Tollefsen et al., 2014; Rovida et al., 2015 |
| Transparency regarding all information sources used, including testing costs, animal numbers, uncertainty; data processing, evaluation target (hazard/dose-response information), endpoint. | Lewis et al., 2007; Ahlers et al., 2008; Kinsner-Ovaskainen et al., 2009; De Wever et al., 2012; OECD, 2016b |
| Flexibility regarding the integration of new information, (e.g. hypothesis-driven approach). | Kinsner-Ovaskainen et al., 2009; Jaworska et al., 2010; Jaworska and Hoffmann, 2010; De Wever et al., 2012 |
| Final decision based on a weight-of-evidence approach. | Jaworska et al., 2010; Jaworska and Hoffmann, 2010; Basketter et al., 2013 |
| Address/document/reduce uncertainty of information generated from individual testing methods and uncertainty of extrapolations to effects in humans. | Lewis et al., 2007; Ahlers et al., 2008; Schaafsma et al., 2009; OECD, 2016b |
| Use of mechanistic information; relate construction of DA to mechanistic understanding of an endpoint (e.g. AOP). | Ankley et al., 2010; Dellarco et al., 2010; Basketter et al., 2013; Landesmann et al., 2013; Vinken, 2013; Rovida et al., 2015; OECD, 2016b |

According to the criteria presented in Table 2.2, information revealed from testing should be reliable, accurate, precise and fit-for-purpose. Uncertainties should be transparently addressed and documented. Furthermore, different cost components need to be acknowledged, in particular welfare losses from misclassifications (Hoffmann and Hartung, 2006), monetary costs of regulatory validation of testing strategies (Schaafsma et al., 2009), animal welfare loss (Hartung et al., 2013; Rovida et al., 2015) and direct testing costs (Krewski et al., 2010; Hartung et al., 2013; Rovida et al., 2015). Jaworska and Hoffmann (2010) indicated that the process of integrating information from toxicity testing should allow for evidence maximisation while considering factors such as costs, animal welfare and test complexity for an optimal selection of information sources (computational, *in chemico* and *in vitro* methods). Although efficiency (Kinsner-Ovaskainen et al., 2009), optimisation (Rovida et al., 2015) or cost effectiveness (Lewis et al., 2007; Hoffmann et al., 2008; Hartung et al., 2013) are mentioned in the toxicological literature, the required methodological approaches for addressing these criteria in the process of developing a DA are not addressed. While the possible trade-offs occurring when developing DAs are acknowledged, still it has largely remained unclear how these trade-offs can be made transparent and how they should be addressed to ensure resource-efficient DAs. Krewski et al. (2010) emphasised the “...the difficulty in simultaneously meeting four objectives ...” when developing DAs, e.g. minimise testing costs, the number of laboratory animals, the time to perform testing methods and simultaneously provide sufficient information (also see (Nordberg et al., 2008; Gabbert and van Ierland, 2010).

2.2.2 An economic perspective to developing optimal defined approaches

The term “*resource efficiency*” is a key economic decision-criterion for guiding the allocation of scarce resources. In the economics’ literature “*resource efficiency*” denotes an allocation of resources that allows achieving a given outcome target with a minimum of resources (Clemen and Reilly, 2001). Toxicological testing of substances aims at generating new information about the hazardous properties of substances. The ultimate goal of toxicity testing is to allow for adopting better-informed decisions upon chemicals’ use. Depending on the toxicological effect of interest (the so-called “endpoint”), toxicological testing requires a variety of resources, in particular appropriate laboratory equipment or computational capacities, manpower, laboratory animals, and time. The challenge is, therefore, to distribute available resources such that a maximum of output – i.e. hazard information – can be achieved, or, to use a dual formulation, that a certain information outcome can be achieved with a minimum of resources (Norlén et al., 2014). Thus, efficient or optimal testing can be

characterised as a process where a maximum of information can be achieved at the lowest cost.

The European chemicals' legislation REACH requires large numbers of industrial chemicals to be tested within defined time-frames (EC, 2006). Acknowledging that existing testing capacities are tight, several studies pointed to the urgent need to structure the testing process more efficiently (Jaworska and Hoffmann, 2010; Krewski et al., 2010; Hartung et al., 2013; Rovida et al., 2015). Corresponding to the efficiency definition provided above this requires specifying key criteria for efficiency evaluations of testing. A basic distinction can be made between (i) conceptual and methodological criteria for balancing gains and costs from testing and (ii) informational criteria (Table 2.3).

Table 2.3: Key criteria for evaluating the resource efficiency of defined approaches (DAs)

| Informational criteria | Possible assessment parameters |
|--|---|
| Specification of information gain/outcome. | <ul style="list-style-type: none"> • Accuracy parameters (e.g. sensitivity/specificity) for characterising the ability of a method to assess hazard/potency classes. • Reliability parameters (intra- and inter-laboratory reproducibility). • Mechanistic information (e.g. mode of action or coverage of key events in the AOP). • Combinations of different parameters (e.g. entropy). |
| Specification of costs. | <ul style="list-style-type: none"> • Direct testing costs (laboratory equipment or computational capacities, animal welfare loss, testing time, labour costs). • Indirect testing costs (i.e. validation costs). |
| Conceptual criteria | Possible methods/approaches |
| Valuation of information gains and costs. | <ul style="list-style-type: none"> • Monetary valuation. • Non-monetary valuation. |
| Purpose of the assessment. | <ul style="list-style-type: none"> • Hazard identification, potency sub-categorisation. |
| Approach to balance information gains and costs. | <ul style="list-style-type: none"> • Qualitative approach (e.g. multi criteria analysis). • Quantitative approach (e.g. cost-effectiveness analysis, cost-benefit analysis). |
| Assessment of uncertainties of parameters assessing information gains and testing costs. | <ul style="list-style-type: none"> • Frequentist statistics' approaches (e.g. calculation of confidence intervals). • Bayesian inference methods. • Approaches for assessing testing method's precision. |
| Stopping rule for testing. | <ul style="list-style-type: none"> • Decision-theoretic approaches (e.g. Value-of-Information analysis). • Mechanistic relevance driven approaches (e.g. the adverse outcome pathway AOP). |

Efficiency evaluations of individual testing methods and DAs require, first of all, specifying information gains and costs. Then we need to select appropriate quantifiable parameters for both components. Testing costs can be further distinguished into direct and indirect costs. Direct costs consist of i.e. (i) laboratory equipment or computational capacities for conducting a testing method which are directly required for conducting a testing method or a combination of methods, (ii) laboratory animal welfare loss (in case of an animal test) and (iii) testing time. Indirect testing costs include, for example, expenditures, resources and time needed for the validation of a (non-animal) testing method, or switching costs for cases where new technologies have to be adopted (Norlén et al., 2014).

First, for quantifying information gains from testing different metrics can be used. For example, a testing method's information outcome can be characterised in terms of its predictive accuracy, describing *“the closeness of agreement between test method results and accepted reference values”* (OECD, 2015a). Common accuracy metrics are sensitivity (i.e. the proportion of hazardous substances correctly classified as hazardous by a testing method) and specificity (i.e. the proportion of non-hazardous substances correctly classified as non-hazardous by the testing method). In addition, information gains from testing can be characterised by a testing method's reliability, denoting a testing method's ability to be reproduced within and between laboratories over time and usually expressed in terms of a testing method's intra- and inter- reproducibility. Finally, information about the coverage of specific key events in the adverse outcome pathway (AOP) of a particular *in vivo* adverse outcome by a specific testing method is important to quantify the informational gains from testing.

Second, efficiency evaluations require defining a mechanism to balance information gains from testing with costs. How to do this depends, ultimately, on how information gains and costs are valued. Basically, two possibilities exist, i.e. monetary and non-monetary valuation. In case of a non-monetary valuation information or cost parameters are expressed in terms of their natural units (e.g. the proportion of positive chemicals correctly classified in case of sensitivity). A monetary assessment requires transferring information or cost parameters into Euro or Dollar values. While direct costs, e.g. expenditures for conducting a test, are usually expressed in monetary terms, a monetary valuation of other cost components (e.g. animal welfare loss) is less common and also often not wanted. Likewise, monetising information is not straightforward. In economics, different approaches have been suggested for quantitatively balancing information gains from testing with costs, some of which have also been applied to the field of toxicity testing (see Section 2.3 for a detailed discussion).

Third, evaluating the efficiency of testing methods, and guiding the development of optimal DA, must account for the uncertainty underlying to information gains and costs. In particular, since any testing method, including the animal test, is a model representation of human or environmental endpoint considered, information outcomes from testing are uncertain. Ideally, if different (non-animal) testing methods are combined into a DA, uncertainty will be reduced throughout the strategy. Again, different options exist for assessing uncertainty of test information and costs. A distinction can be made between frequentist and Bayesian approaches. Frequentist approaches, e.g. the calculation of confidence intervals of predictivity parameters, require the underlying datasets to be of a sufficient size to be meaningful. Bayesian inference methods explicitly account for a decision-makers subjective (prior) beliefs and allow for updating information about uncertainty if new data become available. Finally, given that a DA is a framework that combines individual non-animal testing methods, integrating data from individual testing methods into a DA requires to determine a stopping rule for testing (Gabbert and Weikard, 2013; Leontaridou et al., 2016).

2.3 Methodological approaches for data integration into defined approaches

2.3.1 Qualitative approaches

2.3.1.1 Descriptive weight of evidence (WoE) based approaches

Descriptive Weight of Evidence (WoE) is a qualitative process for evaluating existing information and deciding whether or not further testing is necessary (Grindon et al., 2006b). Testing includes *in silico*, *in chemico*, and *in vitro* methods. Animal tests (*in vivo*) can be used as a “last resort” if evidence from non-animal methods is considered insufficient (Vermeire et al., 2013). For example the overview of the assessment strategy for skin sensitisation (EHCA, 2016), the decision tree described in (Patlewicz et al., 2015) or even the earlier suggestion from (Grindon et al., 2008e) for assessing skin sensitisation guides the collection and evaluation of existing information and decision-making for further testing by means of graphical flowcharts. In a descriptive WoE approach, testing methods are performed either in combination or in a sequential order following flowcharts which indicate the order of tests to be conducted depending on whether information already collected is sufficient. Testing outcomes are usually characterised by their predictive accuracy measures, however, the starting point and the number of testing methods to be performed is decided by expert judgment.

2.3.1.2 Mode of action (MoA) based approaches

The Mode of Action (MoA) concept describes a series of key events which are causally related to a toxic effect. The MoA concept was first applied to assessing potential carcinogens (EPA, 2005) and later it was extended to analysing non-cancer toxic effects (Boobis et al., 2008). It has been suggested as a guiding tool to evaluate existing information and to integrate information from *in-silico*, *in chemico*, and *in-vitro* methods (Lilienblum et al., 2008; Dellarco et al., 2010; Dellarco and Fenner-Crisp, 2012; Simon et al., 2014) into DA frameworks. The MoA approach combines information from testing methods depending on the biological relevance of key events addressed by non-animal testing methods, which are assumed to lead to toxic effects on organs responses (Vonk et al., 2009). The MoA concept has been considered a promising tool to guide toxicity testing because it can be used for prioritising substances and for evaluating information from testing based on their mechanistic-relevance to humans. Therefore, the MoA concept can, first of all, guide toxicity testing with respect to the applicability domain of substances. Second, it can guide the development of testing strategies based on the mechanistic relevance of information derived from non-animal methods regarding the toxic effect on organs.

2.3.1.3 Adverse outcome pathway (AOP) based approaches

The adverse outcome pathway (AOP) concept describes the biological (key) events and their linkages which ultimately lead to the expression of an adverse effect at the level of an organism (OECD, 2012b; OECD, 2012c). Conceptually, the AOP broadens the scope of the MoA by considering adverse effects at an organism- or population- level rather than organ- level addressed by the MoA (Ankley et al., 2010; Vinken, 2013). The AOP has been suggested as a criterion for guiding data collection, and for the organisation and evaluation of relevant information derived from non-animal testing methods (Ankley et al., 2010; Landesmann et al., 2013; Vinken, 2013; Kleinstreuer et al., 2016) in order to develop DAs. The advantage of using the AOP concept as a guiding tool for the construction of testing strategies is its ability to guide the collection and combination of specific pieces of information about key events in a biologically consistent manner (Gocht et al., 2015). AOP-based approaches aim at the full replacement of the animal tests with mechanistically-relevant combinations of non-animal testing methods addressing specific key events in the AOP (Schultz et al., 2016).

2.3.2 Quantitative approaches

2.3.2.1 Machine learning approaches

Machine learning approaches encompass computational algorithms developed to predict hazardous properties of substances and to reduce uncertainties underlying to the assessment of the hazardous properties. During the past years machine learning approaches have been increasingly suggested and applied to identify the types of information, called variables, which have to be combined to draw conclusions about the hazardous properties of substances. Applications of machine learning methods to the construction of DAs are Bayesian networks (BN) (Jaworska et al., 2011; Jaworska et al., 2013), Artificial Neural Networks (ANN) (Hirota et al., 2013; Tsujita-Inoue et al., 2014; Hirota et al., 2015; Tsujita-Inoue et al., 2015), Naïve Bayes Algorithms (NB), Support Vector Machines (SVM) and Classification and Regression Trees (CART) (Matheson, 2015; Asturiol et al., 2016; Kleinstreuer et al., 2016). Using machine learning approaches is considered a suitable approach to optimise DAs because they allow for a quantification of uncertainties at any stage of the testing strategy, and they allow for learning (i.e. updating the assessment) if new information (e.g. about the molecular structure of a substance) is received. Since existing applications of machine learning approaches focus exclusively on the information side of testing these approaches can also be denoted information-theoretic, in contrast to decision theoretic methods which are explained below.

2.3.2.2 Cost analysis approaches

Contrary to machine learning approaches, MoA- and AOP- based approaches, cost analysis approaches allow for balancing informational gains from testing with costs. If information gains and costs of testing can be expressed in monetary terms a cost-benefit analysis (CBA, see Bergstrom and Varian, (2003)) can be applied. Testing methods and DAs can be ranked according to their (expected) net benefits. In absence of monetary values for information gains or cost components, cost effectiveness analysis (CEA) is used (Hurd, 2015). CEA has been proposed and repeatedly been used as a decision-support tool for test selection in a regulatory context (Lave et al., 1988; Omenn, 1995; Bjørner and Keiding, 2004) and for evaluating toxicity testing strategies (Gabbert and van Ierland, 2010). Information outcome may be quantified in terms of a testing method's performance metrics (Gabbert and van Ierland, 2010; Norlén et al., 2014).

2.3.2.1 Decision theoretic approaches

Decision theory approaches aim at quantifying expected net gains of a process (e.g. a policy intervention, a medical treatment), acknowledging that within this process different

decisions/actions can be adopted and that the outcomes of a decision/action are uncertain. Value of Information (VOI) analysis has been proposed as a decision-theoretic tool to prioritise and optimise testing (Lave et al., 1988; Yokota et al., 2004; Gabbert and Weikard, 2013; Leontaridou et al., 2016). Based on expected utility theory, VOI analysis quantifies the expected payoff of any possible decision adopted with and without information from testing (Claxton, 1999). Payoffs can be expressed as welfare gains and losses resulting from decisions upon the use of a substance (e.g. ban or release). In a Bayesian inference framework, VOI analysis allows for incorporating a decision maker's beliefs about the true hazardous properties of a substance and to update beliefs if new information becomes available (Gabbert and Weikard, 2013; Leontaridou et al., 2016). A decision-theoretic framework based on the expected utility maximisation theory has been proposed (Hansson and Rudén, 2007), using frequencies to describe the accuracy metrics of testing methods. Since the expected value of test information is a quantitative measure it allows for ranking testing methods and strategies. In addition VOI analysis can be used to guide the construction of testing strategies because it offers an endogenous rule when testing should stop (i.e. the expected value of test information exceeds testing costs, see Leontaridou et al. (2016)). When a monetisation of test information is not possible, Multi Criteria Decision Analysis (MCDA) can be used. MCDA has been suggested as an approach to integrate evidence from different sources (Linkov et al., 2011; Linkov et al., 2015) for the optimisation on assessments for nanoparticles (Hristozov et al., 2014).

2.4 Evaluating defined approaches addressing skin sensitisation according to resource efficiency criteria

To date, none of the non-animal testing methods can provide sufficient information to fully replace the animal tests used for skin sensitisation hazard identification and potency assessment as standalones (ECHA, 2016). Instead, a combination of *in vitro*, *in chemico* and *in silico* methods has been considered a good way to generate sufficient information and eventually replace *in vivo* tests (Kinsner-Ovaskainen et al., 2009; Casati et al., 2013). During the past years, several testing strategies have been proposed for the assessment of skin sensitisation potential and potency (Jaworska, 2016; OECD, 2016c). Strategies complying with the definition of a DA proposed in (OECD, 2016b) are summarised in (OECD, 2016c) and they are extensively described in the Annex I (OECD, 2016e) of the respective report. Note that these strategies have been also characterised as IATAs (Urbisch et al., 2015b) SEURAT-1 annual report in 2015 (CEFIC, 2015).

DAs for skin sensitisation potential and potency assessment use different conceptual and methodological approaches to integrate information from the individual testing methods. Hence, they are presented in different ways, for example in the form of qualitative flowcharts (Grindon et al., 2008e; Mekenyan et al., 2010; ECHA, 2014a; Patlewicz et al., 2015), probabilistic approaches (machine learning) applying Artificial Neural Networks (Hirota et al., 2013; Tsujita-Inoue et al., 2014; Hirota et al., 2015) or Bayesian Networks (Jaworska and Hoffmann, 2010; Jaworska et al., 2013; Jaworska et al., 2015), and as deterministic approaches based on a “majority vote” decision rule for batteries of testing methods (Bauch et al., 2012; van der Veen et al., 2014a; van der Veen et al., 2014b; Urbisch et al., 2015a) or score-based batteries of testing methods (Ellison et al., 2010; Nukada et al., 2013; Takenouchi et al., 2015). In addition, a regression analysis model (Natsch et al., 2015) and a quantitative model using the toxico-kinetics and toxico-dynamics modelling (MacKay et al., 2013) are used. Based on the criteria defined in Table 2.3 existing DAs for assessing skin sensitisation can be evaluated and compared regarding the resource efficiency of data integration. Table 2.4 shows DAs that were selected as reference examples in (OECD, 2016c).

Table 2.4: Evaluation of defined approaches proposed to assess skin sensitisation hazard or potency according to conceptual and informational criteria

| Name of the DA | | "2 out of 3" ITS | Kao ITS | Kao STS | RIVM STS | Stacking meta model | IDS | BN ITS | ANN ITS | EC-JRC | Global and local regression models | IATA | SARA |
|--|---------------------------------------|---|--|--|--|--|---|--|---|---|--|--|---|
| Type of the DA | | Deterministic | | | Deterministic and probabi- listic | | LLNA and human data | LLNA | LLNA | LLNA | Regression | Decision tree | Exposure based model |
| Informational criteria | Predictive accuracy compared to | LLNA and human data | LLNA | LLNA | LLNA and human data | | LLNA | LLNA | LLNA | LLNA and human data | LLNA and human data | LLNA and GPMT | Human data |
| | Information outcome | Sensitivity Specificity Accuracy | Sensitivity Specificity Accuracy | Sensitivity Specificity Accuracy | 1 st step of strategy: quantita- tive WoE using BN | Sensitivity Specificity Confidence Kappa | Sensitivity Specificity Accuracy | Sensitivity Specificity Accuracy MI metrics | Linear correlation analysis | Sensitivity Specificity Accuracy Balanced accuracy, NPV and ppv | R ² and P values as correlation strength measures | Sensitivity Specificity Accuracy In a WoE judgment | Sensitivity Specificity Balanced accuracy, NPV and ppv |
| | Reliability | Inter- and intra- reproduci- bility of individual methods Interchange- ability of individual methods | Applica- bility domain assessment | Applicability domain assessment | Reliability of individual methods expressed in NPV and ppv | Reprodu- cibility Variability of LLNA as reference | Leave-one- out cross- validation for reliability of DA | Precision assessment using Bayes factors | The goodness- of-fit was evaluated in terms of RMS error | Reproducib ility and robustness of the DA assessed with compari- son of different models | Leave-one- out cross- validation for reliability of DA | n/a | Global sensitivity analysis for exposure, chemical- specific and biological parameters |
| Coverage of KE in AOP | | 1, 2, 3 | 1, 3 | 1, 3 | 1, 2, 3 | 1, 2, 3 | 1, 2, 3 | 1, 2, 3 | 1, 2, 3 | 1 (MIE) | 1, 2 | 1, 2, 3, 4 | 1, 3, 4 |
| Costs Direct and indirect testing costs | | n/a | n/a | n/a | Direct testing costs are reported | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |

Regarding the information outcomes from testing each individual testing method used in a DA is characterised in terms of its predictive accuracy. For determining the predictive accuracy the sets of substances used are well documented and most commonly compared with the reference animal tests such as the LLNA or human data when available. DAs use testing methods for which reliability measures, accounting for inter- and intra-reproducibility of the testing methods are determined, however, they are not always explained in detail. For assessing the reliability of probabilistic DAs cross-validation statistical tools are suggested to cross check the robustness of DAs. For example, for the IDS the reliability of the strategy to predict the LLNA classification, thus the final decision outcome, was checked using the leave-one-out validation (Strickland et al., 2016). For the ANN-ITS, the ability of the model to predict the final decision outcome on the skin sensitisation potential was validated using the 10-fold cross validation approach (Hirota et al., 2013). Probabilistic approaches for the assessment of potency or potential offer the statistical tools to combine information of different parameters such as the BN ITS with the use of Mutual Information (MI) metrics (Jaworska et al., 2010; Jaworska et al., 2013; Jaworska et al., 2015). The mechanistic understanding of the information collected from each source of information is based on identifying key events covered in the skin sensitisation AOP. Generally, it is assumed that covering the first three key events of the AOP is sufficient to draw conclusions on a substance's skin sensitisation potential. This is the case for most DAs presented in Table 2.4 (Jaworska et al., 2013; van der Veen et al., 2014a; Jaworska et al., 2015; Urbisch et al., 2015a). However, some DAs focus on selected key events only. For example, Kao DA (Nukada et al., 2013; Takenouchi et al., 2015) covers the first and third key event, whereas the EC-JRC DA covers the molecular initiating event (MIE) only (Dimitrov et al., 2005; Asturiol et al., 2016), which is considered to determine the final conclusion on the skin sensitisation potential (Asturiol et al., 2016). Given that only animal tests can provide information on the fourth key event (i.e. T-cell proliferation), animal welfare considerations do not allow the direct assessment of the fourth key event. The IATA case (Patlewicz et al., 2014) covers the fourth key event with the use of LLNA when necessary and the SARA case (MacKay et al., 2013) predicts this key event using modelling approaches.

Considering direct and indirect testing costs we observe that only one DA, i.e. the RIVM STS, reported direct testing cost estimates (van der Veen *et al.* 2014). In case of probabilistic DAs (Natsch *et al.* 2015; Matheson *et al.* 2015; Hirota *et al.* 2015; Hirota *et al.* 2013; Tsujita-Inoue *et al.* 2015), it is indicated that due to the saving unnecessary testing will also save costs. Furthermore, the Bayesian network (Jaworska *et al.* 2013) and the "2 out of 3" ITS

(Bauch *et al.* 2012; Urbisch *et al.* 2015a) suggest that a non-animal testing method would be avoided if information collected as a certain step of the strategy would be sufficient to conclude on the skin sensitisation potential. This implies saving additional testing costs. However, cost savings are not quantified and, consequently, are not considered an explicit variable for DA development. As DAs for skin sensitisation usually aim at fully replacing animal testing, only in a few cases (e.g. the non-testing pipeline approach (Patlewicz *et al.*, 2014; Patlewicz *et al.* 2015)) the animal test LLNA is proposed as a “last resort”.

Regarding the conceptual criteria, the purpose of the assessment for which a DA is conducted is to assess either skin sensitisation potential or potency. The DAs presented in Table 2.4 document the information outcomes exclusively in a non-monetary way. Specifically, information outcomes are expressed in terms of predictive accuracy metrics. The final decision, i.e. the conclusion whether testing information is sufficient, or whether further testing is required, requires an exogenous decision rule. For example, the “2 out of 3” ITS (Bauch *et al.* 2012; Urbisch *et al.* 2015b) is based on a majority vote where the decision follows the outcome of two concordant test results. Similarly, the RIVM STS (van der Veen *et al.* 2014) uses as a first step a Bayesian QSAR approach, described in (Rorije *et al.*, 2013), which is followed by tiers of non-animal testing methods. The overall conclusion is also based on a majority vote from test results from sequential steps in the strategy. The Artificial Neural Network (ANN) concept (Hirota *et al.* 2015; Hirota *et al.* 2013; Tsujita-Inoue *et al.* 2015) or the Bayesian networks (Jaworska *et al.*, 2010; Jaworska and Hoffmann, 2010; Jaworska *et al.*, 2013; Jaworska *et al.*, 2015) offer probabilistic approaches to predict skin sensitisation potential or potency using different physicochemical properties and information from non-animal testing methods. The IDS DA (Matheson, 2015; Strickland *et al.*, 2016) uses different machine learning approaches, i.e. ANN, Naïve Bayes algorithm (NB), Classification and regression tree (CART), Linear discriminant analysis (LDA), Logistic regression (LR), Support vector machine (SVM). The SVM is considered to be the most accurate (Matheson, 2015; Strickland *et al.*, 2016) and combines information from non-animal testing methods (i.e. h-CLAT), computational methods and physicochemical properties. The EC-JRC DA uses the Classification Trees (CT) machine learning approach based on *in silico* information to predict skin sensitisation potential (Dimitrov *et al.*, 2005; Asturiol *et al.*, 2016). Machine learning models are developed as information maximising approaches and the stopping rule is not clearly defined. It is rather exogenously set by the information target.

Direct or indirect testing costs are in most cases not reported. None of existing DAs incorporate a mechanism that balances information gains and costs. As shown in Leontaridou

et al. (2016), information from testing cannot be considered regardless costs. In particular, testing costs are decisive for determining when testing should stop (Gabbert and Weikard 2013). Hence, ignoring testing costs means that existing DAs do not provide an endogenous stopping rule to testing. Rather, the stopping rule for deterministic approaches such as the “2 out of 3” ITS (Bauch *et al.* 2012; Urbisch *et al.* 2015b), the RIVM STS (van der Veen *et al.*, 2014a; van der Veen *et al.*, 2014b) and the Kao DA suggestions (Nukada *et al.*, 2013; Takenouchi *et al.*, 2015) is often based on exogenous decision rules such as AOP coverage. In probabilistic DAs such as the BN-ITS (Jaworska *et al.*, 2010; Jaworska *et al.*, 2013; Jaworska *et al.*, 2015), the stopping rule is exogenously determined setting information targets i.e. the prediction of the skin sensitisation potential using the LLNA results as a reference. We argue, however, that the stopping rule should not be pre-defined assuming for example covering most key events in the AOP is a sufficient condition to ensure robust predictions from a testing strategy. Instead, decision-theoretic approaches should define rules under which testing should stop when information no longer contributes to the existing knowledge.

The uncertainty underlying to relevant parameters for information outcomes is assessed in a variety of ways. The BN-ITS proposed by (Jaworska *et al.*, 2010; Jaworska *et al.*, 2013; Jaworska *et al.*, 2015), for example, offers an elaborate uncertainty assessment with regard to predictive accuracy of each individual method, and the precision, being the ability of a method to produce concordant results from repeated testing. Uncertainty is based on Bayesian inference and mutual information theory. In case of deterministic approaches, the majority vote is frequently applied to test results without explicitly assessing uncertainties. Individual testing methods reproducibility, interchangeability and reliability is assessed for methods used in, for example, the “2 out of 3” ITS (Bauch *et al.*, 2012; Natsch *et al.*, 2013; Urbisch *et al.*, 2015a) and the RIVM STS (van der Veen *et al.*, 2014a; van der Veen *et al.*, 2014b).

2.5 Conclusions and recommendations

This chapter reviews the state-of-the-art regarding the development DAs denoting approaches to integrate information from different testing and computational methods to determine the hazardous properties or risks of substances. According to the OECD, DAs are defined *“rule-based [approaches] and can either be used on their own ... or considered together with other sources of information in the context of IATA”* (OECD, 2016b). We identified criteria that were suggested in the toxicological literature as normative principles for constructing DAs. One criterion that has frequently been suggested is cost-efficiency. We defined key criteria for evaluating the resource-efficiency of DAs, and explain economic approaches that have been used or suggested for improving resource efficiency of DAs. Using these criteria we

evaluated DAs suggested for the assessment of skin sensitisation potential and potency, and presented in a recent OECD guidance document (OECD, 2016b).

Based on our evaluation we can conclude that none of the existing DAs integrate both information and cost parameters. Instead, DAs predominately focus on maximising information, while the reduction of testing costs is often mentioned to be an important aim of DAs. Still, direct or indirect costs components were not systematically incorporated in the construction of testing strategies. Furthermore, the uncertainty of test information, which can originate from different sources (Worth and Cronin, 2001b; Kolle et al., 2013; Hoffmann, 2015; Leontaridou et al., 2017a; Leontaridou et al., 2017b) are assessed only in some of the existing DAs. Basically, ignoring direct and indirect testing costs implies that the resource efficiency of DAs cannot be evaluated. In particular, DAs discussed in this paper do not allow for identifying the trade-offs between generating information from testing, and the costs for resources required to attain new information from testing.

As a consequence, it remains unclear whether DAs indeed allow for optimising toxicological testing compared to animal tests, if “optimising” is interpreted in terms of economic resource efficiency. Moreover, suggested DAs for skin sensitisation testing lack an endogenous stopping rule. Evaluating the resource efficiency of DAs requires, first, to document both information outcomes and costs of testing. In addition, information gains must be balanced with costs. This can be achieved by means of integrating cost information in the construction method of a DA. In particular, machine learning approaches offer the possibility to weighing costs and information gains while accounting for uncertainties of both parameters at any stage of the DA. In a recent paper by Leontaridou et al. (2016) Bayesian VOI analysis has been applied to optimising DAs for skin sensitisation testing. Their approach showed that DAs are more resource efficient compared to the animal test LLNA. Alternatively, the resource-efficiency of DAs can be evaluated ex post, i.e. after the strategy was developed, using cost analysis methods, i.e. CBA or CEA. Both methods have been widely used for efficiency assessments of, for example, medical treatments, where conceptual challenges to identify the best performing alternative are very similar to toxicity testing (Claxton, 1999; Cunningham, 2001; Bergstrom and Varian, 2003; Claxton et al., 2004). It is important to note that optimal, i.e. resource-efficient, testing does not require that all relevant parameters are monetised. However, further research is required to assess direct and indirect costs of toxicity testing to ensure that cost information can be integrated in the construction of DAs. We believe that this is a prerequisite for developing optimal testing approaches which ensure valid safety assessments chemicals without animal testing, and low cost.

3 Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian Value-of-Information analysis²

This chapter offers a Bayesian Value-of-Information (VOI) analysis for guiding the development of non-animal testing strategies, balancing information gains from testing with the expected social gains and costs from the adoption of regulatory decisions. Testing is assumed to have value, if and only if, the information revealed from testing triggers a welfare-improving decision on the use (or non-use) of a substance. As an illustration, our VOI model is applied to a set of five individual non-animal testing methods used for skin sensitisation hazard assessment, seven battery combinations of these methods, and 236 sequential 2-test and 3-test strategies. Their expected values are quantified and compared to the expected value of the local lymph node assay (LLNA) as the animal method. We find that battery and sequential combinations of non-animal testing methods reveal a significantly higher expected value than the LLNA. This holds for the entire range of prior beliefs. Furthermore, our results illustrate that the testing strategy with the highest expected value does not necessarily have to follow the order of key events in the sensitisation adverse outcome pathway (AOP).

² Chapter 3 is published as: Leontaridou M., Gabbert S., et al., (2016). Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian Value-of-Information analysis. *ATLA Alternatives to Laboratory Animals* vol. 44(3), pp. 255-269.

3.1 Introduction

Skin sensitisation denotes an immunological response that results in allergic contact dermatitis (ACD) after repeated exposure to a sensitising substance (Kimber et al., 2002; ECHA, 2014a). Besides being a key endpoint for safety evaluations of cosmetic ingredients, skin sensitisation testing is mandatory for all substances produced or marketed in volumes larger than 1 tonne per year under the European REACH legislation (EC, 2003a; EC, 2006). Although REACH does not prescribe a strict replacement of animal testing, it bases testing requirements on the paradigm of using *in vivo* testing only “as a last resort” ((EC, 2006), Article 25). In addition to REACH, the Cosmetics Regulation enforced a phasing-out of toxicity testing in animals by establishing a marketing ban for finished cosmetic products with ingredients tested in animals, which came into force in March 2013 (EC, 2003a; Hartung, 2010a). This fostered the development of new, animal-free testing methods for assessing skin sensitisation hazard.

Several non-animal testing methods for predicting skin sensitisation have been developed (see (Reisinger et al., 2015), for a detailed overview). Of these, one *in chemico* method and two *in vitro* methods were formally validated, i.e. the Direct Peptide Reactivity Assay (DPRA; (Gerberick et al., 2007)), the human Cell-Line Activation Test (h-CLAT; (Sakaguchi et al., 2006)) and the ARE-Nrf2 luciferase method covered by KeratinoSensTM (Emter et al., 2010). The last-named testing method is also covered by LuSens (Ramirez et al., 2014), for which validation is still pending. Organisation for Economic Co-operation and Development (OECD) Test Guidelines (TGs) have been adopted for the DPRA (OECD TG 442C; (OECD, 2015a)) and the ARE-Nrf2 luciferase method (OECD TG 442D; (OECD, 2015b)). Nevertheless, none of the available non-animal methods satisfy the requirements for being accepted as an individual replacement of the animal-based method. The main reasons are that non-animal methods usually cover only selected steps of the entire adverse outcome pathway (AOP) for inducing an allergic reaction in the skin, and existing non-animal methods are unable to deliver information on skin sensitisation potency (Goebel et al., 2012; Worth et al., 2014).

The lack of full replacement options has shifted attention to possibilities for *combining in vitro*, *in silico* and *in chemico* methods into batteries and sequential (integrated) testing strategies (Rovida et al., 2015). Ultimately, the development of integrated approaches for testing and assessment (IATAs) should solve fundamental problems, in particular, which testing methods to select and how to combine different methods in a strategy (Patlewicz et al., 2014); the latter issue also includes the problem of defining the optimum number of steps required. In recent studies on integrated testing strategies for skin sensitisation, the selection

and sequential ordering of non-animal testing methods has been guided by the AOP for skin sensitisation (Patlewicz et al., 2014; van der Veen et al., 2014a). This implies that covering all consecutive events in an AOP delivers more-reliable information about a substance's properties. Furthermore, the few studies offering a quantitative evaluation of the performance of non-animal testing strategies focus exclusively on information gains from testing, being either expressed as discrete predictivity estimates for skin sensitisation hazard classification (van der Veen et al., 2014a; Vinken et al., 2014), or as probability predictions for substances belonging to a specific potency class (Jaworska et al., 2013).

This chapter argues that the criteria for guiding the construction of non-animal testing strategies, and their evaluation as replacements of animal tests should be based on an approach that accounts for testing costs and the social gains and losses from possible regulatory decisions, rather than solely on information gains from testing. This is motivated by three arguments. First, despite the progress in AOP development during recent years (Vinken, 2013; Tollefsen et al., 2014; Patlewicz et al., 2015), for many adverse outcomes the knowledge of the AOP is still rudimentary. For endpoints for which the AOP is well characterised, there is no plausible reason why a testing strategy should necessarily cover all the key events. On the contrary, it could be preferable to start a testing strategy with the most informative method. Second, none of the available methods, including animal tests, deliver perfect information. Thus, irrespective of the information metric used, test information is uncertain. In several studies, test information has been interpreted from a Bayesian perspective as conditional probabilities that update a decision-maker's beliefs about the hypothesis that a substance has a specific property. As shown in (Gabbert and Weikard, 2013), the construction of non-animal testing strategies with regard to information gains alone is insufficient, because it lacks an intrinsic rule of when testing should stop. Third, testing is costly (Koch and Ashford, 2006). Hence, assuming that the ultimate goal of testing is to inform regulatory decision-making, which, in turn, aims at improving social welfare, an evaluation of non-animal testing methods and testing strategies must balance information gains against costs.

The objectives of Chapter 3 are twofold. Our first objective is to introduce a decision-theoretic Value-of-Information (VOI) approach for developing and evaluating non-animal testing strategies. We assume that testing has a 'value', if, and only if, the information revealed from testing triggers a welfare-improving decision on the use (or non-use) of a substance, compared to decision-making in the absence of additional information from testing. Thus, in contrast with information-theoretic approaches such as Bayesian Networks, Hidden Markov

or quantitative Weight-of-Evidence approaches (Rorije et al., 2013; van der Veen et al., 2014a; Luechtefeld et al., 2015; Rovida et al., 2015), VOI analysis explicitly considers expected social gains and costs (called “payoffs”) from any possible decision on the use of a substance, while accounting for the uncertainty of test information. Quantifying the VOI provides a tool which guides the choice and sequencing of methods in a testing strategy. By comparing the VOI of different testing methods and testing strategies, the tool offers insight into the fundamental question of whether, and under what conditions, the VOI of a non-animal testing strategy outperforms the VOI of an animal test.

The second objective is to illustrate the features of our model by applying it to the case of skin sensitisation hazard assessment. This complements and expands the information-theoretic literature on the development of non-animal testing strategies for skin sensitisation (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Jaworska et al., 2013) by also incorporating societal benefits and costs of testing into the analysis. The VOI is calculated for a set of individual non-animal testing methods, including *in vitro*, *in silico* and *in chemico* methods, seven battery combinations, 62 two-test and 174 three test sequential combinations of these non-animal testing methods (Urbisch et al., 2015a). Their VOIs are compared to the VOI of the local lymph node assay (LLNA) as the animal test (Mehling et al., 2012; Basketter et al., 2014; Urbisch et al., 2015a).

3.2 Method: A decision-theoretic approach to assessing the value of testing

VOI analysis is a decision-analytic method that calculates expected gains and losses from gathering additional information. It has been widely applied to problems of decision-making surrounded by uncertainty in many different domains, such as medical diagnosis and healthcare decision-making, environmental technology assessment, environmental pollution management and the prioritisation of regulatory strategies (Claxton, 1999; Claxton et al., 2004; Yokota et al., 2004; Yokota and Thompson, 2004). Several studies applied VOI analysis to the problem of chemical risk management (Lave et al., 1988; Yokota et al., 2004; Gabbert and Weikard, 2010; Gabbert and Weikard, 2013) of which two studies offer applications to sequential combinations of tests, focusing on carcinogenicity and mutagenicity hazard assessments (Yokota and Thompson, 2004; Gabbert and Weikard, 2013). However, Yokota and Thompson (2004) pre-defined the selection and ordering of tests. Also, the sets of non-animal testing methods addressed by Gabbert et al. (2013) and by Yokota and Thompson (2004) were confined to *in vitro* methods only, and the animal test is combined with non-animal testing methods into a sequential testing strategy. A comprehensive evaluation of different non-animal testing methods and their combinations, assessing their potential to

replace the animal test, has not been conducted so far. In this paper, the value of collecting additional information is assessed from a social welfare perspective, where gains and losses of both producers and consumers are considered. Depending on exposure, a substance can lead to the manifestation of an adverse effect on human health or on ecosystems (an “endpoint”; (van Leeuwen et al., 2007; Wallace and Ernest, 2012)). The simplest approach is to identify an adverse effect with a binary “hazardous/non-hazardous effects” assessment (van der Schouw et al., 1995; Hoffmann and Hartung, 2005; Rovida et al., 2015). In the following, we will use $\tau = 1$ for denoting that a substance has the potential to cause a hazardous effect, and $\tau = 0$ if it does not. Depending on a decision-maker’s beliefs about the state of the substance, they must decide whether the substance can be used. For simplicity, we assume that the set of regulatory decisions contains only two options: “release” (i.e. allow access to the market) and “ban” (i.e. prohibit access to the market). Depending on τ and the set of regulatory actions, the use of a substance can have benefits, but also costs, to society.

Assuming a competitive market, the social benefits of releasing a non-hazardous substance are the sum of the producer’s expected marketing benefits B_p (being the difference between revenue and costs; (Bergstrom and Varian, 2003), and expected intermediate consumer benefits B_c (intermediate consumers are companies who use a substance as ingredient for their products). The release of a hazardous substance, in contrast, can cause health and environmental damages, which are costly for society. Social costs, D , comprise health and environmental damage costs from the use of a substance. Thus, benefits and costs of the use of chemicals are incurred by different economic factors. If a substance is released, expected social damage costs must be subtracted from expected social benefits. If a substance is banned, we assume zero social benefits, irrespective of the true hazardous properties of the substance. Clearly, expected payoffs are substance- specific. Table 3.1 summarises expected payoffs for the possible states of a substance and the set of regulatory actions, i.e. the action space.

Table 3.1: Payoffs from substances’ use

| Action space | State of substance | |
|-----------------|--------------------|-------------|
| | $\tau = 1$ | $\tau = 0$ |
| | | |
| Ban | 0 | 0 |
| Release | $B_p + B_c - D$ | $B_p + B_c$ |

$\tau = 1$: Substance is hazardous;

$\tau = 0$: Substance is non-hazardous;

B_p = Marketing benefits of chemical producer;

B_c = Marketing benefits of intermediate consumer;

D = expected health damage costs.

Prior to testing, a decision-maker's beliefs about the state of a substance depend on available information about τ . This could be based on information about a substance's structure (e.g. taken from an OECD profiler screening), earlier studies, evidence of the prevalence of toxic health effects, or expert judgement. In the absence of any information, prior beliefs can be completely subjective (Held and Bové, 2014). Prior beliefs that a substance is hazardous are denoted p_0 , and $(1 - p_0)$ if the substance is believed to be non-hazardous. Based on prior beliefs, a substance should be released if, and only if,

$$p_0(B_P + B_C - D) + (1 - p_0)(B_P + B_C) > 0 \quad (3.1)$$

According to Eq.3.1, a chemical should be marketed if the probability-weighted sum of payoffs is positive. Assuming that decision-makers aim at maximising expected payoffs, the value of decision-making under uncertainty, i.e. without information from testing, is:

$$V_0 = \max [0; (p_0(B_P + B_C - D) + (1 - p_0)(B_P + B_C))] \quad (3.2)$$

That is, a ban will be preferred if the expected payoff (0, left expression in square brackets in Eq. 3.2) exceeds the expected payoff from releasing the substance. The latter is the sum of probability-weighted payoffs for marketing a hazardous and a non-hazardous substance, respectively (right expression in square brackets in Eq. 3.2). Testing reveals additional information, which reduces uncertainty about the true status of the substance, provided that the methods used are reliable and relevant. For hazard identification, the continuous dose-response curve resulting from testing is often dichotomised into a binary hazardous/non-hazardous classification (van der Schouw et al., 1995; Hoffmann and Hartung, 2005). We define t^+ as a positive test outcome, indicating that a substance is hazardous. A negative test outcome is denoted t^- . So far, neither the animal test nor the available non-animal testing methods provide perfect information. Thus, testing may reduce, though not fully resolve, uncertainty. The predictive capacity of a non-animal testing method has usually been expressed by means of a 2×2 contingency tables (Cooper et al., 1979), which shows the proportion of correct (sensitivity s , and specificity r) and false classifications, i.e. the false negative rate $(1 - s)$ and false positive rate $(1 - r)$, based on a pre-defined training set of substances with known properties (Table 3.2; (Bauch et al., 2012; Urbisch et al., 2015a)). From a Bayesian perspective, the proportion of correct and false classifications can be interpreted as conditional probabilities of seeing a positive or negative test result, given that

the chemical is hazardous or non-hazardous, $p(t|\tau)$ (Jaworska et al., 2010; Jaworska et al., 2013; Rorije et al., 2013).

Table 3.2: Conditional probabilities $p(t|\tau)$ of seeing a testing result t , given the true state of the substance τ

| | | State of the substance | |
|--------------|-------|--|--|
| | | $\tau = 1$ | $\tau = 0$ |
| Test outcome | | s | $1 - r$ |
| | t^+ | Probability of a true positive outcome (Sensitivity) | Probability of a false positive outcome |
| | | $1 - s$ | r |
| | t^- | Probability of a false negative outcome | Probability of true negative outcome (Specificity) |

$\tau = 1$: Substance is hazardous;

$\tau = 0$: Substance is non-hazardous;

t^+ : Test outcome is positive;

t^- : Test outcome is negative.

Additional information may change a decision maker's beliefs about the state of the substance. Using Bayes' theorem (Eq. 3.3), which is the standard approach for probabilistic information update and learning (Howson and Urbach, 1991), the decision-maker's prior beliefs p_0 can be revised into posterior beliefs on the state of the substance $p(\tau|t)$, given the outcomes from testing (Table 3.3):

$$p(\tau|t) = \frac{p_0 p(t|\tau)}{p_0 p(t|\tau) + (1 - p_0)(1 - p(t|\tau))}. \quad (3.3)$$

Table 3.3: Posterior probability $p(\tau|t)$ of a substance being hazardous or non-hazardous after seeing evidence from testing

| | | State of the substance | |
|--------------|-------|--|---|
| | | $\tau = 1$ | $\tau = 0$ |
| Test outcome | | $p(\tau t)$ | |
| | t^+ | $p^+ = \frac{p_0 s}{p_0 s + (1 - p_0)(1 - r)}$ | $1 - p^+ = \frac{(1 - p_0)(1 - r)}{(1 - p_0)(1 - r) + p_0 s}$ |
| | | $p(\tau t)$ | |
| | t^- | $p^- = \frac{p_0(1 - s)}{p_0(1 - s) + (1 - p_0)r}$ | $1 - p^- = \frac{(1 - p_0)r}{(1 - p_0)r + p_0(1 - s)}$ |

$\tau = 1$: Substance is hazardous; $\tau = 0$: Substance is non-hazardous;

t^+ : Test outcome is positive; t^- : Test outcome is negative;

p_0 : Prior probability that a substance is hazardous; $(1 - p_0)$: Prior probability that the substance is non-hazardous;

s : Probability of a true positive outcome (sensitivity); $(1 - s)$: Probability of a false negative outcome;

r : Probability of a true negative outcome (specificity); $(1 - r)$: Probability of a false positive outcome;

p^+ : Posterior probability that the substance is hazardous after seeing a positive test outcome;

$1 - p^+$: Posterior probability that the substance is hazardous after seeing a negative test outcome;

p^- : Posterior probability that the substance is non-hazardous after seeing a positive test outcome;

$1 - p^-$: Posterior probability that the substance is non-hazardous after seeing a negative test outcome.

In the specific case that prior beliefs are equivalent to the prevalence of toxic health effects, the posterior probability of a substance being hazardous or non-hazardous coincides with a testing method's positive and negative predictive value, respectively. The expected value of taking an optimal action under posterior beliefs is, then,

$$V_i^+ \equiv \max [0; (p^+(B_P + B_C - D) + (1 - p^+)(B_P + B_C))], \quad (3.4)$$

in case of a positive test result, and

$$V_i^- \equiv \max [0; (p^-(B_P + B_C - D) + (1 - p^-)(B_P + B_C))] \quad (3.5)$$

in case of a negative test result.

The expected value of performing test i is the weighted sum of V_i^+ and V_i^- , with the probability of seeing a positive and a negative test outcome (Pr^+ , Pr^-) being the weights:

$$V_i = Pr^+ V_i^+ + Pr^- V_i^-, \quad (3.6)$$

where

$$Pr^+ = p_0 s + (1 - p_0)(1 - r) \quad (3.7a)$$

denotes the probability of seeing a positive test outcome, and

$$Pr^- = p_0(1 - s) + (1 - p_0)(r) \quad (3.7b)$$

denotes the probability of seeing a negative test outcome.

It follows from Eq. 3.2 and Eq.3.6 that a test i should be performed if the expected value from an optimal decision with additional information from testing, V_i , exceeds the expected value of an optimal decision without information from testing, V_0 . Thus, the expected value of test information ($EVTI_i$) is

$$EVTI_i = V_i - V_0. \quad (3.8)$$

Since testing is costly, a test i should be performed if, and only if, its expected value exceeds testing costs:

$$EVTI_i - k_i \geq 0, \quad (3.9)$$

with k_i denoting monetary testing costs (Norlén et al., 2014). Note that Eq. 3.9 provides the rule for testing to be stopped.

The higher the $EVTI_i$ net of testing costs, the higher the expected social benefits arising from the use of a substance, given additional information from test i . For a set of non-animal testing methods, the one revealing the highest $EVTI_i$ net of costs, will be preferred. A decision-maker will stop testing as soon as the $EVTI_i$ net of testing costs becomes negative (Eq. 3.9). This also holds if the decision maker is a social planner (e.g. a member of a regulatory agency). Although regulators do not bear testing costs, we assume that they adopt a social welfare perspective, because they must consider all types of costs and benefits from the use of a chemical. Bayesian inference allows the calculation of information gains in terms of posterior beliefs about a substance being hazardous or non-hazardous at any stage of the sequence. Thus, Eq. 3.9 offers a quantitative measure for comparing and ranking individual non-animal testing methods, as well as sequential or battery combinations of these methods. A key feature of sequential testing strategies is that the decision of whether or not to continue testing is conditional on information gains at earlier stages in the sequence. This involves the possibility to save tests and, consequently, costs.

3.3 Application: Optimised testing strategies for assessing skin sensitisation hazard of cosmetic ingredients

The applicability of the Bayesian VOI model is illustrated for the case of skin sensitisation hazard assessment of cosmetic ingredients. Given the important role of this endpoint in various regulatory frameworks (Basketter et al., 2012; Luechtefeld et al., 2015), several studies have proposed probabilistic approaches for data integration and the development of hypothesis-driven testing strategies. These studies focus on information gains from testing (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Jaworska et al., 2013; van der Veen et al., 2014a). This paper complements existing approaches by adopting a social welfare perspective, where information gains from testing are balanced with societal benefits and costs from decisions on the use (or non-use) of a substance.

The set of testing methods consists of five non-animal testing methods, seven battery combinations of these methods, and the LLNA as the animal test (Basketter et al., 2012; Mehling et al., 2012; Urbisch et al., 2015a). The non-animal testing methods include the DPRA

(Gerberick et al., 2007), the OECD profiler toolbox v3.2 (denoted hereinafter as “OECD Toolbox”; (OECD, 2012d)), the ARE-Nrf2 luciferase method covered by KeratinoSens™ (Emter et al., 2010) and LuSens (Ramirez et al., 2014), and the h-CLAT (Sakaguchi et al., 2006). The predictive capacities of all the non-animal testing methods and the LLNA were evaluated on human data (subset B in (Urbisch et al., 2015a), derived mainly from (Basketter et al., 2014)), using a sample of 114 substances (Table 3.4). In the skin sensitisation AOP, the DPRA captures the first key event (protein binding), KeratinoSens™ and LuSens the second key event (epithelial responses), and the h-CLAT captures the third key event (dendritic cell activation; (OECD, 2012a)). Note that the cellular non-animal testing methods inherently cover cysteine reactivity and thus mechanistically overlap with the DPRA (Natsch et al., 2013). Finally, we included the OECD Toolbox in the analysis (OECD, 2012d) by using the protein-binding profilers based on OECD and OASIS algorithms, the “auto-oxidation profiler” and the “skin metabolism profiler” (Urbisch et al., 2015a). The analysis was performed for individual non-animal testing methods, selected battery combinations of these methods, and all possible two-test and three test sequences that can be constructed from the five non-animal testing methods in our set. This revealed a total number of 236 sequential testing strategies for assessing skin sensitisation hazard; note that we did not consider combinations of methods where an individual method would be applied repeatedly.

Table 3.4: Predictive capacity and testing costs for individual methods and battery combinations

| Testing methods | Sensitivity [%] | Specificity [%] | Testing costs ^a [Euro/substance] |
|---|-----------------|-----------------|--|
| LLNA ^b | 91 | 64 | 5,500 |
| OECD toolbox ^b | 89 | 64 | 500 |
| DPRA ^c | 84 | 84 | 3,000 |
| KeratinoSens™ ^c | 82 | 84 | 4,000 |
| LuSens ^c | 78 | 79 | 4,000 |
| hCLAT | 89 | 64 | 4,500 |
| DPRA + LuSens ^b | 93 | 100 | 7,000 |
| DPRA + KeratinoSens™ ^d | 100 | 82 | 7,000 |
| DPRA + hCLAT ^b | 94 | 88 | 7,500 |
| LuSens + hCLAT ^b | 96 | 91 | 8,500 |
| KeratinoSens™ + hCLAT ^b | 93 | 92 | 8,500 |
| DPRA + KeratinoSens™ + hCLAT ^d | 90 | 90 | 11,500 |
| DPRA + LuSens + hCLAT ^d | 90 | 89 | 11,500 |

Test batteries are indicated by ‘+’;

^a Estimated average costs 2015 (unpublished);

^b BASF (2015), personal communication;

^c (Urbisch et al., 2015a);

^d Suggested as an AOP-based testing strategy for skin sensitisation hazard assessment, also called the “2-out-of-3ITS”.

Table 3.4 shows that battery combinations of non-animal testing methods have a higher sensitivity/specificity than the LLNA, but are more expensive. The decision-analytic VOI model accounts for the possible trade-offs between information and costs. Quantifying social benefits and costs, in particular, expected health and environmental damage costs caused by the release of a hazardous substance, has been a major challenge in empirical VOI applications (Ennever et al., 1990; Omenn, 1995; Yokota and Thompson, 2004; Gabbert and Weikard, 2013).

A substance that is known to be a skin sensitiser can cause allergic skin reactions (allergic contact dermatitis [ACD]; (Kimber et al., 2002; Park and Zippin, 2014)). Given uncertain test outcomes, there is a risk of harm from ACD, even if the substance has been classified a non-sensitiser. These “costs of making errors” are the monetary health damage costs to society (Table 3.1) — in the following, denoted D_{ACD} . Focusing on non-occupational exposure, health damage costs of an individual suffering from ACD (i.e. D_{ACDin}) consist of direct costs for primary and secondary healthcare (i.e. treatments provided by general practitioners and dermatologists), and of indirect costs (for example, loss of productivity and quality of life). Estimates of health costs caused by ACD provided in the literature vary depending on the severity of ACD (usually expressed in terms of the categories ‘mild’, ‘moderate’ or ‘severe’) and the population group considered (Verboom et al., 2002; Ricci et al., 2006; Halvarsson and Loden, 2007; Stein et al., 2007; Witt et al., 2009; Sætterstrøm et al., 2014). More importantly, empirical cost assessments usually cover only a fraction of overall costs caused by ACD, because several cost components (e.g. loss of quality of life) are difficult to quantify and monetarise. To exemplify the features of our model, we used the mean estimate of health damage costs to an individual adult, as published in (Sætterstrøm et al., 2014), of 973 Euros per person and per year. This estimate includes the direct costs of medical treatment and costs of productivity loss. This value can, of course, differ considerably across individuals within a country and across countries. Furthermore, assessments of health damage costs vary according to the endpoint (see, for example, (Yokota et al., 2004; Yokota and Thompson, 2004)). An alternative approach to estimating the welfare loss of people suffering from ACD would be to determine their “willingness to pay” (WTP) to avoid skin allergies caused by cosmetics. Conducting a revealed preference study is, however, time and resource consuming and was beyond the scope of this paper. WTP estimates provided in a recent ECHA study (ECHA, 2014b) considered, in addition to cosmetic ingredients, a mixed set of allergens, and were based on direct health costs only.

Calculating expected health damage costs to society requires health costs to individuals to be aggregated. Clearly, individuals suffering from ACD caused by specific cosmetic ingredients account for only a fraction of the total population affected by ACD. This is expressed by a substance's sensitisation prevalence ρ , this being the proportion of people in a sample showing positive human patch test reactions to the application of a specific substance (Schnuch et al., 1997; Natsch et al., 2013). Multiplying individual health damage costs (D_{ACDin}) by the prevalence- weighted population ($\rho * N$) provides an estimate of the expected health damage costs (D_{ACD}) caused by ACD. Since our study focuses on Europe, N denotes inhabitants within the European Union (EU28) (EC - Eurostat).

$$D_{ACD} = D_{ACDin} \rho N. \quad (3.10)$$

Prior to the testing of an uncharacterised substance, the 'true' prevalence will be unknown. If a substance belongs to a certain group of contact allergens (e.g. disinfectants, dyes, fragrances, preservatives), a mean prevalence for this group can be used (Schnuch et al., 2011; Leiva-Salinas et al., 2014). If no information is available, a mean prevalence value for contact allergens can be applied (Schnuch et al., 2011).

To calibrate the model, we use the case of methylisothiazolinone. It is used for the formation of Kathon CG, which is a preservative that has been widely used in cosmetic products and is known to be a sensitiser (Uter et al., 2012). Values for the prevalence of sensitisation to MI and Kathon CG have been reported in the literature (Schnuch et al., 2011; Schnuch et al., 2012), and vary, depending on population sample size and composition, between 1.2% and 4.2%. To illustrate the impact of prevalence estimates on the $EVTI_i$, and on the ranking of testing methods and testing strategies included in our analysis, we calculated health damage costs for four different sensitisation prevalence estimates, therefore capturing a range between optimistic and conservative estimates (Table 3.5).

Table 3.5: Prevalence estimates and prevalence-weighted health damage costs caused by ACD within the European Union (EU28)

| Sensitisation prevalence [%] | Source | Prevalence calculation | D_{ACDin} Euro/person and year |
|------------------------------|------------------------|--|-------------------------------------|
| 0.7 | (Schnuch et al., 2012) | Sensitisation prevalence of N-Isopropyl-N'-phenyl-p-phenylenediamine (IPPD), typically low prevalence observed for contact allergens | 7 |
| 2.8 | (Uter et al., 2013) | Prevalence derived from the MOAHLFA index for Kathon CG corrected by the fraction of people suffering from atopic contact dermatitis | 27 |
| 3.8 | (Uter et al., 2013) | Prevalence for Kathon CG from a sample of 28,922 patch test results conducted in the period of 2009-2012 | 37 |
| 15 | (Schnuch et al., 2012) | Sensitisation prevalence of Nickel Sulphate, highest prevalence observed for contact allergens | 146 |

Finally, applying the VOI model requires the determination of expected producer and consumer benefits from a substance's release (B_P and B_C ; Table 3.1). Since data on substance specific revenue and production costs are not available, industry's marketing benefits of Kathon CG were approximated by assuming profits to be 15% of marketing revenues in 2014. For Kathon CG, we used an average price of 1.635 Euros/kg, calculated from monthly prices for the solution of Kathon CG between January 2013 and May 2015 (Zaubä). The quantity of Kathon CG marketed in the EU was assumed to be 1000 tonnes per year, which is the upper limit of the REACH tonnage band (100–1000 tonnes/year; (EC, 2006) Article 12).

Intermediate consumer benefits, B_C , refer to the marketing gains of companies using a cosmetic ingredient — in our case Kathon CG — in their products. Empirical data on the profit share of single cosmetic ingredients are not available. Therefore, it was not possible to quantify B_C directly. Assuming that, prior to testing and in the absence of adequate toxicity information, a decision-maker will likely decide not to release a substance, we approximated B_C as the threshold benefit at which a ban would still just be the optimal action. For uninformative prior beliefs ($p_0 = 0.5$) and re-arranging Eq. 3.1, the threshold consumer benefit is

$$\widehat{B}_C = (p_0 D_{ACD} - B_P) - 1. \quad (3.11)$$

3.4 Results

Applying the VOI model to the set of five non-animal testing methods and the LLNA, seven battery combinations and 236 sequential testing strategies of non-animal methods, and considering four sensitisation prevalence estimates, we obtained a rank list of non-animal methods and testing strategies according to their net $EVTI_i$ of testing costs (see Eq. 3.1). For ease of presentation, the discussion of results is confined to outcomes for sensitisation prevalences of 2.8% and 15%. Furthermore, we show $EVTI_i$ results for two selected prior beliefs, $p_0 = 0.2$ and $p_0 = 0.7$. The first value approximates the prevalence of sensitisers in REACH registration dossiers (Thyssen et al., 2007), the latter denotes beliefs which are slightly more conservative than the percentage of sensitisers in the human data set presented in (Urbisch et al., 2015a). A complete rank list of all individual testing methods, battery combinations and sequential testing strategies for all prior beliefs and prevalence estimates is documented in a supplementary file (available from the website, www.atla.org.uk). The file also includes a numerical example of calculating the $EVTI_i$ net of testing costs for an individual testing method.

Generally, we find that a higher sensitisation prevalence - all other parameters being the same - increases the expected value of testing. The reason is that a higher prevalence causes expected social health damage costs, D_{ACD} , to increase. Consequently, testing becomes more relevant, because it reduces uncertainty and, thus, the probability of the erroneous release of a substance. This is reflected by a higher $EVTI_i$.

Table 3.6 lists the non-animal testing strategies taking the first ten positions: a) for $p_0 = 0.2$, indicating weak prior beliefs that a substance is a sensitiser; and b) for $p_0 = 0.7$, indicating moderate prior beliefs that a substance is a sensitiser. A "+" between two non-animal testing methods denotes a testing battery, and arrows indicate a sequential combination of testing methods. In addition to documenting the testing methods in the ranking, Table 3.6 presents numerical values of their $EVTI_i$ net of testing costs and the incremental difference of the $EVTI_i$ net of testing cost between subsequent rank positions.

We find that individual testing methods and the LLNA are positioned at the end of the ranking. This holds for the entire range of prior beliefs and across all sensitisation prevalence estimates (see the supplementary file for a complete ranking of all testing methods and testing strategies). For very low and very high prior beliefs ($p_0 = 0.1$; $p_0 = 0.9$), the $EVTI_i$ net of testing costs of the LLNA and the individual testing methods equals zero. Thus, testing has no value because expected information from the testing will not change the decision made on the use of the substance, compared to the situation without additional testing information. For

prior beliefs $p_0 \leq 0.4$, the battery DPRA + LuSens reveals the highest $EVTI_i$ net of testing costs (see Table 3.6 and the supplementary file). For prior beliefs $p_0 \geq 0.4$, sequential testing strategies perform best. More specifically, for prior beliefs in the range $0.4 \leq p_0 \leq 0.8$ the sequence consisting of DPRA + LuSens, the OECD Toolbox, and KeratinoSensTM + h-CLAT takes the first five rank positions. This holds for all prevalence scenarios considered. As shown in Table 3.6 for $p_0 = 0.7$, differences in the $EVTI_i$ net of costs across these rank positions are relatively small. The reason is that the $EVTI_i$ difference between two sequences that differ only in the order of testing methods cannot exceed the difference in testing cost between the cheapest and the most expensive test in the sequence. In contrast, the difference of the $EVTI_i$ net of testing costs between strategies with a different composition of non-animal testing methods is usually much larger (see, for example, the strategies on rank position 7 and 8 for $p_0 = 0.7$, and the strategies on rank position 5 and 6 for $p_0 = 0.7$ in Table 3.6).

Table 3.6: Top ten testing strategies according to the $(EVTI_i - k_i)$ assuming a sensitisation prevalence of 2.8% and 15%

| p_0 | Top-10 testing strategies | Sensitisation prevalence 2.8% | | Sensitisation Prevalence 15% | |
|-------|--|-------------------------------|-------------|------------------------------|--------------|
| | | Increment ^a | | Increment ^a | |
| | | $EVTI_i - k_i$ | [Euro] | $EVTI_i - k_i$ | [Euro] |
| 0.2 | 1. DPRA + LuSens | 1,285,515,010 | 0 | 6,886,718,050 | 0 |
| | 2. OECD Toolbox → DPRA + LuSens | 1,285,514,510 | -500 | 6,886,717,550 | -500 |
| | 3. KeratinoSens TM → DPRA + LuSens | 1,285,511,010 | -3,500 | 6,886,714,050 | -3,500 |
| | 4. ^b OECD Toolbox → KeratinoSens TM → DPRA + LuSens | 1,285,510,510 | -500 | 6,886,713,550 | -500 |
| | 4. ^b h-CLAT → DPRA + LuSens | 1,285,510,510 | 0 | 6,886,713,550 | 0 |
| | 5. OECD Toolbox → hCLAT → DPRA + LuSens | 1,285,510,010 | -500 | 6,886,713,050 | -500 |
| | 6. ^b h-CLAT → KeratinoSens TM → DPRA + LuSens | 1,285,506,510 | -3,500 | 6,886,709,550 | -3,500 |
| | 6. ^b KeratinoSens TM + h-CLAT → DPRA + LuSens → OECD Toolbox | 1,285,506,510 | 0 | 6,886,709,550 | 0 |
| | 7. OECD Toolbox → KeratinoSens TM + hCLAT → DPRA + LuSens | 1,285,506,010 | -500 | 6,886,709,050 | -500 |
| | 8. DPRA + KeratinoSens TM → LuSens + hCLAT | 1,237,408,681 | -48,097,328 | 6,629,018,318 | -257,690,732 |
| 0.7 | 9. DPRA + KeratinoSens TM → OECD Toolbox → LuSens + h-CLAT | 1,237,408,509 | -172 | 6,629,018,146 | -172 |
| | 10. LuSens + h-CLAT → DPRA + KeratinoSens TM | 1,237,408,257 | -252 | 6,629,017,894 | -252 |
| | 1. DPRA + LuSens → OECD Toolbox → KeratinoSens TM + hCLAT | 1,955,348,612 | 0 | 10,475,118,725 | 0 |
| | 2. OECD Toolbox → DPRA + LuSens → KeratinoSens TM + hCLAT | 1,955,348,286 | -326 | 10,475,118,399 | -326 |
| | 3. DPRA + LuSens → KeratinoSens TM + hCLAT → OECD Toolbox | 1,955,347,074 | -1,213 | 10,475,117,187 | -1,213 |
| | 4. OECD Toolbox → KeratinoSens TM + hCLAT → DPRA + LuSens | 1,955,345,191 | -1,883 | 10,475,115,304 | -1,883 |
| | 5. KeratinoSens TM + hCLAT → DPRA + LuSens → OECD Toolbox | 1,955,341,540 | -3,651 | 10,475,111,653 | -3,651 |
| | 6. DPRA + KeratinoSens TM → OECD Toolbox → LuSens + hCLAT | 1,896,272,102 | -59,069,438 | 10,158,636,821 | -316,474,832 |
| | 7. OECD Toolbox → DPRA + KeratinoSens TM → LuSens + hCLAT | 1,896,271,979 | -123 | 10,158,636,698 | -123 |
| | 8. DPRA + KeratinoSens TM → LuSens + hCLAT → OECD Toolbox | 1,896,266,980 | -4,999 | 10,158,631,699 | -4,999 |
| | 9. OECD Toolbox → LuSens + hCLAT → DPRA + KeratinoSens TM | 1,896,265,672 | -1,308 | 10,158,630,391 | -1,308 |
| | 10. LuSens + hCLAT → DPRA + KeratinoSens TM → OECD Toolbox | 1,896,264,889 | -783 | 10,158,629,608 | -783 |

Estimates of the predictive capacities of individual testing methods were extracted from (Urbisch et al., 2015a) and are based on human data.

^a The increment denotes the difference of the $EVTI_i - k_i$ to the preceding testing method or testing strategy

^b Testing strategies have the same $EVTI_i - k_i$ and, therefore, share the same rank position.

The better performance of sequential testing strategies compared to batteries of non-animal testing methods can be explained by their potential to reduce the number of tests needed, and, hence, to reduce testing costs. This can be demonstrated by looking into the belief updating process of each branch in a sequence. Figure 3.1 illustrates the sequential structure of the first ranked obtained for $p_0 = 0.7$, being DPRA + LuSens \rightarrow OECD Toolbox \rightarrow KeratinoSensTM + h-CLAT. If DPRA + LuSens would reveal a positive result (DPRA + LuSens: t_1^+), conducting the OECD Toolbox at the second stage and KeratinoSensTM + h-CLAT at the third stage of the sequence would not sufficiently shift posterior beliefs in order to change the optimal action from “Ban” to “Release”. As a consequence, the expected value of the OECD Toolbox and KeratinoSensTM + h-CLAT would be zero, irrespective of whether their outcome would be positive or negative.

If the use of the battery DPRA + LuSens would reveal a negative result (t_1^-), the $EVTI_i$ net of testing costs of the OECD Toolbox would be negative. This can be explained as follows: At the second stage of the sequence, the expected value of the OECD Toolbox is equivalent to the expected value of the battery DPRA + LuSens at the first stage ($V_1^- = V_2|V_1^- = 4970$ million Euros). Hence, one would expect the $EVTI_i$ after the second stage of the sequence to be zero. However, conditional on seeing a negative result of the OECD Toolbox, a positive outcome of KeratinoSensTM + h-CLAT at the third stage of the sequence, would shift the action from “Release” to “Ban”. This is reflected by a positive $EVTI_i$ after the third stage of the sequence ($EVTI_3|V_{1,2}^{-+} = 1455$ million Euros). Conducting KeratinoSensTM + h-CLAT at the third stage implies that the second testing stage must have been conducted as well, and the costs of the second testing method would be incurred. Therefore, the $EVTI_i$ of the OECD Toolbox is negative and equivalent to its costs (–500 Euros). Thus, whereas in a battery all testing methods are conducted (therefore incurring testing costs of all methods), in the case of a sequence, whether a testing method of a follow up stage will be conducted or not depends on the outcomes observed at previous stages.

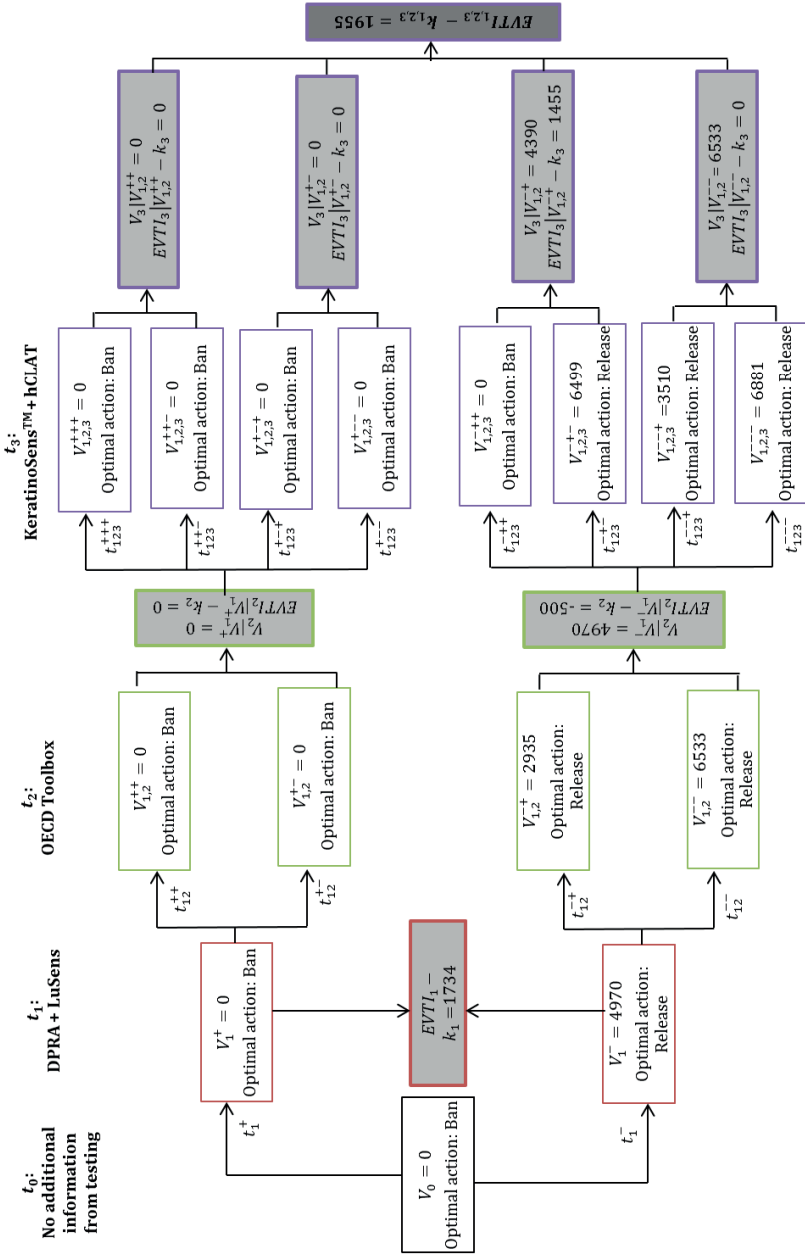


Figure 3.1: Testing strategy revealing the highest $EVTI_1$ net of costs for prior belief $p_0=0.7$, and for a sensitisation prevalence of 2.8%

Non-shaded boxes show the expected values and the expected optimal action for a particular outcome of a testing method. Light grey shaded boxes show the probability-weighted expected values and the $EVTI_i$ net of costs at the first, second and third stage of the sequence, respectively. The $EVTI_i$ net of costs of the entire sequence is the dark grey shaded box. All values are expressed in million Euros.

As illustrated in Table 3.7, the $EVTI_i$ of testing strategies depends on a decision-maker's prior beliefs. This emphasises the need to carefully evaluate all available information prior to testing, e.g. from screening methods or expert consultation, in order to determine meaningful prior probabilities. Comparing numerical $EVTI_i$ results of the first ranked strategies at different prior beliefs with that of the LLNA underlines that — given the assumptions and data discussed in the Method section — the $EVTI_i$ of battery and sequential combinations of non-animal testing methods is significantly higher. To put our results in the context of AOP-based prediction models for skin sensitisation hazard assessment suggested in the literature (29, 45), we compared the $EVTI_i$ net of costs of first-ranked strategies with that of the “2 out of 3” ITS, where the classification for skin sensitisation is based on congruent results of at least two of three non-animal testing methods (Urbisch et al., 2015a). Here, we considered the DPRA, KeratinoSens™ and the h-CLAT (see also Table 3.4). Our results demonstrate that the $EVTI_i$ net of testing costs of this strategy, though being higher than that of the LLNA, is lower than the $EVTI_i$ net of testing costs of the first-ranked strategies in Table 3.6 for all prior beliefs. Hence, if used as a battery (as in (Urbisch et al., 2015a)), the “2 out of 3” ITS, despite its high joint sensitivity and specificity, is outperformed by sequential combinations of non-animal testing methods.

Table 3.7: $EVTI_i - k_i$ (Million Euros) of the first- ranked testing strategy for different prior beliefs (p_0), and for a sensitisation prevalence of 2.8%^a

| p_0 | First-ranked testing strategy | $EVTI_i - k_i$ of first-ranked strategy | $EVTI_i - k_i$ of LLNA | $EVTI_i - k_i$ of “2 out of 3” ITS ^b |
|-------|---|---|------------------------|---|
| 0.1 | DPRA + LuSens DPRA + LuSens → OECD Toolbox → KeratinoSens™ + hCLAT | 642,754,005 | 0 | 0 |
| 0.2 | | 1,285,515,010 | 0 | 691,129,366 |
| 0.3 | | 1,928,276,014 | 145,134,082 | 1,382,270,231 |
| 0.4 | | 2,611,781,558 | 1,022,882,981 | 2,073,411,096 |
| 0.5 | | 3,314,491,729 | 1,900,631,879 | 2,764,551,961 |
| 0.6 | | 2,634,920,170 | 1,396,099,048 | 2,073,411,095 |
| 0.7 | | 1,955,348,612 | 891,566,216 | 1,382,270,230 |
| 0.8 | | 1,275,777,053 | 387,033,384 | 691,129,365 |
| 0.9 | DPRA + KeratinoSens™ → OECD Toolbox → LuSens + hCLAT | 830,318,758 | 0 | 0 |

^a Results of the first-ranked strategy are compared to the ($EVTI_i - k_i$) of the LLNA and of the “2-out-of-3” ITS, as suggested in (Bauch et al., 2012; Urbisch et al., 2015a)

^b This approach consists of the DPRA, KeratinoSens™ and the hCLAT. The conclusion on a substance's hazardous properties is based on a majority vote; the h-CLAT will be conducted only if the first two prediction methods are in disagreement (Bauch et al., 2012; Urbisch et al., 2015a).

Finally, our results illustrate that testing strategies do not necessarily have to follow the order of key events in an AOP, nor do all key events have to be covered. Looking into the list of methods taking the ten top positions in the ranking (Table 3.6), we observe that testing strategies for assessing skin sensitisation hazard do not have to start with methods covering protein activation, nor do all events in the AOP for skin sensitisation have to be covered (Figure 3.2). A possible explanation for this finding is that combinations of testing methods reduce the variation in predictions of either method, which increases the overall predictive capacity of the testing strategy. This underlines that key events in an AOP, if sufficiently known, cannot be understood as a construction rule for testing. Instead, with which non-animal testing method (or combination of methods) a testing strategy should start, and in which order to conduct testing methods, depend on the interplay between prior beliefs, the predictive capacity of methods, and the specification of payoffs and testing costs. The strength of the VOI framework is to offer a transparent analysis of whether, and under what conditions, a non-animal testing strategy provides sufficient and adequate hazard information for optimal decision-making without full coverage of all AOP events.

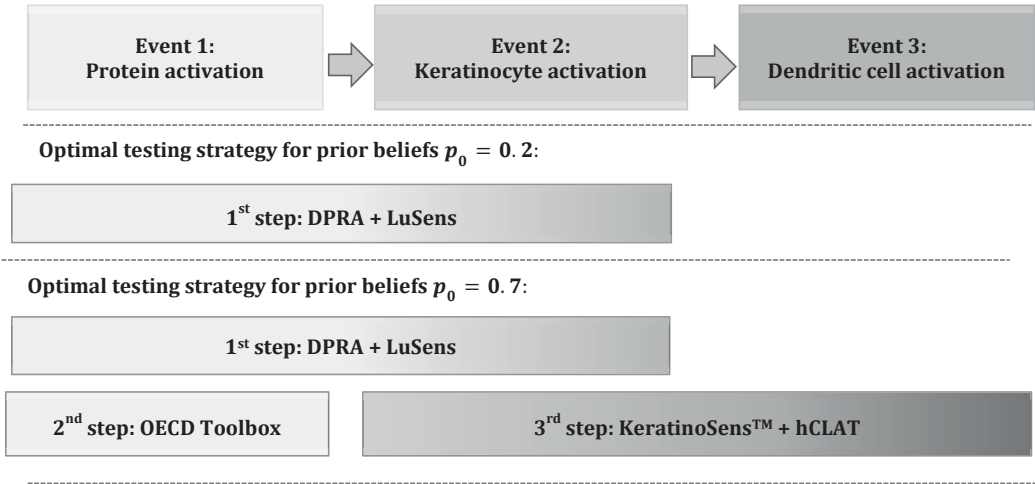


Figure 3.2: Schematic representation of key events in the skin sensitisation AOP covered by the first-ranked testing strategies for prior beliefs of $p_0=0.2$ and $p_0=0.7$, and a skin sensitisation prevalence of 2.8%

3.5 Discussion and conclusions

Bayesian VOI analysis is a decision-theoretic tool for assessing a testing method's potential for improving decision-making on the use of a substance. Testing has a value, if expected social net gains from an optimal decision with additional evidence outweigh expected net gains from decision making without such evidence. Hence, a testing method, or any combination of methods, should be performed if the $EVTI_i$ is positive, and if it exceeds testing costs. The $EVTI_i$ net of testing costs can be quantified for individual testing methods, and battery and sequential combinations of methods. VOI analysis can therefore be used for comparing and ranking different testing options. In addition, the VOI framework guides the construction of sequential testing strategies, because it permits the determination of which order of methods reveals the highest $EVTI_i$ net of testing costs, and when testing should stop.

This chapter complements the existing literature on developing and optimising integrated testing strategies. During recent years, several studies have addressed the challenge of constructing non-animal testing strategies for skin sensitisation hazard classification and potency assessment. These studies have focused on the maximisation of informational gains from sequential testing strategies and batteries (Jaworska and Hoffmann, 2010; Jaworska et al., 2011; Jaworska et al., 2013; van der Veen et al., 2014a). The role of social benefits and costs for developing testing strategies, and how to balance information gains against theoretic costs, have not been addressed. In order to complement existing information-theoretic approaches, our paper applies decision-theoretic VOI analysis which integrates information gains and costs of chemical use. Moreover, it adopts a social welfare perspective.

As an illustration, we applied the model to the problem of skin sensitisation hazard assessment. We quantified the $EVTI_i$ net of testing costs for a set of validated or pre-validated non-animal methods (including the DPRA, the OECD Toolbox, the ARE-Nrf2 luciferase method covered by KeratinoSens™ and LuSens, and the h-CLAT), seven battery combinations of these methods, and 236 sequential 2-test and 3-test strategies composed of these methods. Their $EVTI_i$ net of testing costs was compared with that of the animal test LLNA. Social benefits and costs from the release of cosmetic ingredients to the market were calculated by estimating industry's marketing gains and health damage costs caused by ACD. Clearly, the numerical estimates used in our study and, in particular, estimates of marketing benefits and health damage costs, may vary across countries and between chemicals.

The impact of variations in health damage costs was analysed by means of a sensitivity analysis. Moreover, the predictive capacity of non-animal testing methods and their combinations depends on the number and the selection of substances in the training set.

Therefore, $EVTI_i$ estimates can vary if training sets are composed differently. In addition, conditional dependence between individual non-animal testing methods may affect their joint sensitivity and specificity when used in combination. Accounting for this uncertainty, e.g. by using categorical data analysis (Zaubas), might change the predictive capacity of testing methods and, following on from this, their $EVTI_i$. Since the main purpose of our study was to illustrate the features of decision-theoretic VOI analysis, a detailed assessment of these uncertainties is beyond the scope of this paper, but remains an interesting aspect for further research.

Our results warrant a number of interesting conclusions. Firstly, the value of combinations of non-animal testing methods outweighs that of the LLNA. This is a robust result, because it holds for both battery and sequential combinations, for the entire range of prior beliefs, and for all the sensitisation prevalence estimates considered. Furthermore, sequential combinations of batteries can outperform individual battery combinations of non-animal methods. In our case study, this became already apparent at relatively low prior beliefs. One explanation for this result is the higher predictive capacity of battery combinations in comparison to individual non-animal methods, which reduces the probability of adopting erroneous decisions. More importantly, sequential testing strategies offer the possibility to save on testing methods, and, therefore, on testing costs. The reason is that, depending on the outcomes at previous stages in the sequence, follow-up testing methods would not be conducted in any case. This “cost-saving potential” of sequential testing strategies results in a higher $EVTI_i$ compared to battery combinations of testing methods.

Secondly, for given predictive capacities of non-animal methods, payoff estimates and testing costs, the optimal order of tests is highly sensitive to a decision-maker’s prior beliefs. Hence, unlike the approach suggested in REACH and recent studies on the testing of skin sensitisation (Bauch et al., 2012; Basketter et al., 2012; van der Veen et al., 2014a), our results underline that there cannot be a pre-defined ‘best approach’ to testing. In contrast, determining the optimal testing strategy must consider the trade-offs between expected gains and costs, while accounting for the uncertainties inherent in all parameters.

Finally, our results illustrate that full coverage of all key events in the skin sensitisation AOP is neither a necessary nor a sufficient condition for combinations of non-animal testing methods revealing a higher $EVTI_i$ net of testing costs than the LLNA. However, this only holds if the key events of an AOP are known. Therefore, further research should address possible extensions of the decision-theoretic VOI model to permit the identification of the optimal combination of methods for endpoints where knowledge about the AOP is still incomplete.

Likewise, there is no reason to assume that a non-animal testing strategy must necessarily cover each key event in the sensitisation AOP just once. Since information outcomes of individual non-animal testing methods are uncertain, there is a probability that the prediction is wrong. If different non-animal testing methods are combined into a battery, uncertainties of individual methods can compensate each other, which reduces the probability of erroneous predictions. As our results show, it can therefore be optimal to combine batteries of non-animal methods into a sequence, even though this might repeatedly address the same key event in the AOP. In other words: From a decision-analytic perspective, the AOP concept can neither be understood as a rule for the combination nor for the order of non-animal testing methods in a sequence.

Clearly, the numerical results of our analysis depend on model assumptions and the quality of the data used. Furthermore, we focused on non-occupational exposure for the calculation of sensitisation prevalence. The inclusion of occupational exposure may increase estimated health damage costs, which changes the $EVTI_i$ of the testing methods. Since this applies symmetrically to all testing methods and testing strategies, the ranking of methods will remain unchanged. The calculation of expected payoffs was based on simplified assumptions due to lacking data, for example about substance-specific benefits for intermediate consumers. The model can, however, be used straightforwardly, to investigate the impact of increasing or decreasing social benefits and costs on the $EVTI_i$ net of testing costs — for example, if in-house marketing data for substance groups or particular substances are available, or if testing costs change over time. Finally, although our case study was characterised by high health damage costs in relation to marketing benefits and testing costs, this may not hold for all endpoints. Hence, expanding the analysis to other endpoints and non-animal testing methods will offer further insights into the potential of non-animal testing methods to replace animal testing.

4 The borderline range of prediction models for skin sensitisation potential assessment: Quantification and implications for evaluating non-animal testing methods' precision³

Testing methods to assess the skin sensitisation potential of a substance usually use threshold criteria to dichotomise continuous experimental read-outs into “yes/no” conclusions. The threshold criteria are prescribed in the respective OECD test guidelines and the conclusion is used for regulatory hazard assessment, i.e. classification and labelling of the substance. Due to biological and technical variability we can identify a borderline range (BR) around the classification threshold within which test results are non-conclusive. We quantify the BR of the prediction models of the non-animal testing methods DPRA, LuSens and h-CLAT. The borderline ranges were between $\pm 10\%$ and $\pm 30\%$ of the respective testing methods' thresholds. We find that of the 199, 79 and 40 substances tested 20, 5 and 8 (10%, 6% and 20%) were borderline with the DPRA, LuSens and the h-CLAT, respectively. If the results of individual non-animal test methods are combined into integrated testing strategies (ITS), borderline test results of individual tests can affect the overall assessment of the skin sensitisation potential of the testing strategy. This was analysed for the “2 out of 3” ITS: Four out of 40 substances (10%) were actually borderline. This compares to six out of the 22 (27%) performance standard substances of the LLNA. Based on our findings we propose expanding the standard binary classification of substances into “positive/negative” or “hazardous/non-hazardous” by adding a “borderline” or “non-conclusive” alert for cases where test results fall within the borderline range.

³ Chapter 4 is an earlier version of the manuscript: Leontaridou M., Urbisch D., Kolle S.N., Ott K., Mulliner D.M., Gabbert S. and Landsiedel R., (2017). The borderline range of toxicological methods: Quantification and implications for evaluating precision (ALTEX in press doi: 10.14573/altex.1606271.).

4.1 Introduction

Skin sensitisers are substances that can lead to an allergic response following skin contact (UNECE, 2011). An individual will be sensitised upon first contact. Subsequent contact can then provoke allergic contact dermatitis (ACD). It is estimated that ACD affects about 20% of the European and North American population at least once in their lifetime, although there is considerable variation of skin sensitisation prevalence between different age-sex groups (Thyssen et al., 2007). Data on skin sensitisation potential have to be provided for all substances produced or manufactured above one tonne per year under the European chemicals legislation REACH, and for classification and labelling of substances under the European CLP regulation (ECHA, 2016). The assessment of a substance's skin sensitisation potential has been traditionally based on data derived from animal tests such as the guinea pig based tests described in OECD TG no. 406 (OECD, 1992) or the murine local lymph node assay (LLNA) described in OECD TG no. 429 (OECD, 2002; OECD, 2010). However, animal welfare concerns, and the regulatory enforcement e.g. by the Cosmetics Regulation (EC, 2009) and the REACH legislation (EC, 2006) have driven efforts to move away from animal to non-animal testing. A number of non-animal testing methods have been developed (Mehling et al., 2012; Reisinger et al., 2015), two of which, namely the Direct Peptide Reactivity Assay (DPRA) (Gerberick et al., 2004; Gerberick et al., 2007) and the antioxidant response element - nuclear factor erythroid 2 (ARE-Nrf2) luciferase testing methods covered by KeratinoSens™ (Natsch et al., 2011), have been validated by the European Centre for Validation of Alternative Methods (ECVAM; Italy) and are described in the OECD TG no. 442C and no. 442D (OECD, 2015a; OECD, 2015b). LuSens (Ramirez et al., 2014; Ramirez et al., 2016) also covers the ARE-Nrf2 luciferase testing method and is currently undergoing validation. Another non-animal testing method, the human cell line activation test (h-CLAT) (Ashikaga et al., 2006; Sakaguchi et al., 2006; Ashikaga et al., 2010; Sakaguchi et al., 2010) has recently been validated by ECVAM and is described in OECD TG no. 442E (OECD, 2016d). The sequential structure of molecular and cellular mechanisms causing ACD is represented by the “adverse outcome pathway” (AOP) for skin sensitisation, consisting of eleven causally linked steps, four of which were defined to be essential and specific “key events” (OECD, 2012b; OECD, 2012c). The DPRA, the ARE-Nrf2 testing methods and the h-CLAT cover the first three key events of the skin sensitisation AOP.

For hazard classification purposes, i.e. for assessing skin sensitisation potential, continuous data obtained from animal tests or from non-animal testing methods are

dichotomised into binary “positive/negative” information (van der Schouw et al., 1995; Hoffmann and Hartung, 2005). The prediction models used for the DPRA, LuSens and the h-CLAT are described in OECD TG no. 442C (OECD, 2015a), Ramirez et al. (2014 and 2016), and in the OECD TG no. 442E (OECD, 2016d), respectively. Based on the threshold for classification a testing method’s accuracy, i.e. the percentage of true positive and true negative classifications, can be determined (see for example (Cooper et al., 1979) and (Yerushalmy, 1947)).

The experimental data obtained from a testing method are, however, subject to biological and technical variability. As a consequence, repeated testing may result in discordant classification results. This impacts the precision of a testing method, defined as the ability of a testing method to deliver concordant results in repeated applications. The problem of intra- and inter-assay variability of *in vitro* methods has been observed earlier (see Hothorn (2002 and 2003)). Luechtefeld et al. (2016) pointed to a limited intra-assay reproducibility of skin sensitisation potential and potency data.

This chapter focuses on the intra-assay variability of testing methods for skin sensitisation potential assessment. Specifically, we analyse limitations with regard to the reproducibility of results when continuous dose-response data are transformed into “toxic/ non-toxic” outcomes. Kolle et al. (2013), Hoffmann (2015), Dumont et al., (2016), and Dimitrov et al. (2016) analysed the intra-assay variability of the LLNA. Kolle et al. (2013), showed that outcomes of repeated testing of a substance are not always concordant. Specifically, for those substances for which the estimated concentration (EC₃) leads to a simulation (SI) index value which was relatively close to the threshold for classification (i.e. SI = 3; Kolle et al., 2013), different classifications: Positive or negative for skin sensitisation can result. Kolle et al. (2013), defined a range around the classification threshold of the LLNA, within which discordant outcomes can be expected, by determining coefficients of variation based on individual animal data. This range is called “borderline range” (BR) (Kolle et al., 2013), or “grey zone” (Dimitrov et al., 2016). The percentage of substances falling into the BR of a testing method’s prediction model can be used as a measure of the, i.e. the intra-assay variability, i.e. the testing method’s limited precision.

Analyses of the BR for non-animal testing methods used for skin sensitisation potential assessment have not been conducted so far. Furthermore, a comparative evaluation of the precision of non-animal testing methods and the LLNA has not become available. The aim of this chapter is, therefore, to fill this gap by examining the impact of technical and biological variability on the precision of selected non-animal testing methods for skin sensitisation

potential assessment. Moreover, we compare the precision of the non-animal testing methods with that of the animal test LLNA. For this purpose the BR was quantified for the non-animal testing methods DPRA, LuSens, h-CLAT and the LLNA, based on results revealed from a large number of experiments (Appendix A). The approach for quantifying the BR and the decision rules for detecting borderline substances in experimental samples are explained in Section 4.2. Results from quantifying the BR for each individual testing method are presented in Section 4.3.1. Borderline substances detected in the experimental samples of individual testing methods are shown in Section 4.3.2. Finally, Section 4.3.3 shows borderline substances for the “2 out of 3” ITS. Section 4.4 discusses implications from considering the BR in non-animal testing methods’ prediction models and the “2 out of 3” ITS, respectively. Section 4.5 concludes.

4.2 Materials and methods

4.2.1 Testing methods

The three non-animal testing methods DPRA, LuSens, and h-CLAT were developed to address the three key events of the AOP in order to assess a substance’s skin sensitisation potential. We compared our findings to those of the LLNA as *in vivo* reference test to evaluate the precision of these methods. The samples used for quantifying the BR contained 42 substances in case of the DPRA, 26 substances in case of LuSens, 13 substances in case of the h-CLAT, and 22 substances in case of the LLNA, respectively. The BR was quantified using results from a large number of runs of each testing method. Information about the samples used for determining the BR for each non-animal testing method and the LLNA, the number of runs conducted and the substance concentrations used in the experiments is provided in Appendix A, Tables A1-A4. Where substance names could not be provided due to data confidentiality substances were numbered consecutively.

The experimental samples to which the BR concept was applied in order to detect borderline substances consists of 199 substances in case of the DPRA, 79 in case of LuSens, 40 in case of the h-CLAT, and 22 substances in case of the LLNA; see Bauch et al. (2012) and Urbisch et al. (2015a, 2016). The composition of these samples is presented in Appendix B, Tables B1-B5.

4.2.1.1 Local lymph node assay

The Local Lymph Node Assay (LLNA) is the “first choice” animal test for the assessment of skin sensitisation potential (Kimber et al., 1994). It is described in OECD TG 429, which was first published in 2002 (OECD, 2002) and updated in 2010 (OECD, 2010). In the LLNA, the proliferation of lymphocytes in auricular draining lymph nodes induced by substances is quantified by comparing the mean proliferation in each test group to the mean proliferation in the vehicle treated control group. The ratio of the mean proliferation in each treated group to that in the concurrent vehicle control group, termed the Stimulation Index (SI), is determined. The classification threshold of the LLNA is $SI = 3$. If $SI > 3$ a substance is classified a skin sensitiser.

4.2.1.2 Direct peptide reactivity assay

The Direct Peptide Reactivity Assay (DPRA) was developed by (Gerberick et al., 2004; Gerberick et al., 2007). The DPRA has been formally validated and the OECD Testing Guideline TG 442C (OECD, 2015a) was adopted in 2015. In the DPRA, depletions of two model peptides containing a cysteine- or lysine- residue as a reactive nucleophilic centre are measured after incubation with a test substance. The classification threshold of the DPRA is the mean depletion of 6.38% of the two peptides compared to the depletion in the reference controls (OECD, 2015a). If the mean lysine- and cysteine- peptide depletion is above this threshold, a test substance is considered to be peptide reactive. According to OECD TG 442C the DPRA can be used, together with complementary information, to discriminate sensitisers and non-sensitisers. Depending on the regulatory framework a positive result of the DPRA can serve as standalone information for classifying substances into Category 1 for skin sensitisation. However, as emphasised in the ECHA Guidance on information requirements and Chemical Safety Assessment Chapter R.4a (ECHA, 2016) the DPRA should not be used in isolation for identifying a skin sensitiser or non-sensitiser.

4.2.1.3 ARE-Nrf2 luciferase method

The ARE-Nrf2 luciferase method utilises the gene induction regulated by the antioxidant response element (ARE) in transgenic human keratinocyte cell lines. The OECD Test Guideline TG 442D (OECD, 2015b) was adopted in 2015. The ARE-Nrf2 luciferase method is covered by KeratinoSens™ (Natsch et al., 2011) and LuSens (Ramirez et al., 2014). In this study the LuSens assay is used. In ARE-Nrf2 luciferase methods the keratinocyte activating potential is determined by measuring luciferase induction after treatment with a test substance treatment relative to concurrent vehicle controls. A statistically significant fold induction (FI) of the

Borderline range

luciferase activity above 1.50 is considered to indicate a keratinocyte activating potential of a test substance. The classification threshold for LuSens is FI = 1.50, above which a substance is considered to have a keratinocyte activating potential. Similar to the DPRA, LuSens is not considered suitable for classifying substances as skin sensitisers or non-sensitisers when used in isolation (ECHA, 2016).

4.2.1.4 Human cell line activation test

The human Cell Line Activation Test (h-CLAT) (Ashikaga et al., 2006; Sakaguchi et al., 2006; Ashikaga et al., 2010; Sakaguchi et al., 2010) determines the dendritic cell activating potential by measuring the induction of the expression of the cell surface markers CD54 and CD86 after treatment with a test substance relative to concurrent vehicle controls in immortalised human monocytic leukemia THP-1 cells as a surrogate of DCs. As indicated in the OECD testing guideline TG 442E (OECD, 2016d) a two-fold induction of the CD54 expression and/or 1.50 fold induction of CD86 expression at relative cell viabilities of at least 50% is considered to indicate a dendritic cell activating potential of a test substance. The classification thresholds for h-CLAT are $CD54FI = 1.50$ and $CD86FI = 2.00$. Like for the DPRA and LuSens, the method only addresses a specific key event of the skin sensitisation AOP. Consequently, it should not be used in isolation for classifying skin sensitisation potential (ECHA, 2016).

4.2.1.5 The “2 out of 3” ITS for characterising skin sensitisation potential

The “2 out of 3” ITS (Bauch et al., 2012; Urbisch et al., 2015a; OECD, 2016b; OECD, 2016c) is an integrated testing strategy for the assessment of skin sensitisation potential. According to this approach, 2 out of 3 concordant test results using the DPRA, the ARE-NrF2 luciferase method, and the h-CLAT determine the prediction. The ARE-NrF2 luciferase method can be covered by LuSens or KeratinoSens™. The “2 out of 3” ITS addresses the first three consecutive key events of the AOP for skin sensitisation and it is a selected case study for integrated approaches to testing and assessment (IATA) (Urbisch et al., 2015a). Applying the BR concept to the “2 out of 3” ITS provides a measure for evaluating the performance of this specific DA case.

4.2.2 Approach to quantify the borderline range (BR)

The first step of the variability assessment was to quantify the BR. The BR denotes the area around the classification threshold for which a testing method's prediction model may deliver discordant results. For each non-animal testing method considered, and for the animal

test LLNA, we derived the BR (Eq. 4.1) from the pooled standard deviation SD_p (Eq. 4.2) of a testing method's results (Appendix A), pooled across substances i and concentrations j used (i.e. the dose in case of the LLNA). For the specific case that test results are normally distributed, and the classification threshold is at the mean of the distribution, the BR covers ~68% of the probability mass under the distribution of all test results. Note that the BR approach used in this paper goes beyond Kolle et al. (2013), who calculated the BR only for the LLNA and based on individual animal data. We use the pooled standard deviation SD_p (Eq. 4.2) to define the BR (Eq. 4.1) around a prediction model's classification threshold T . Using the notation shown in Table 4.1 the BR is calculated as follows:

$$BR = \{T - SD_p, T + SD_p\}. \quad (4.1)$$

The pooled standard deviation of experimental results, retrieved from testing different substances and concentrations, is calculated as follows:

$$SD_p = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^{k_i} (r_{i,j}-1) \sigma_{i,j}^2}{\sum_{i=1}^n \sum_{j=1}^{k_i} (r_{i,j}-1)}}, \quad (4.2)$$

where $\sigma_{i,j}^2$ is the variance of the testing methods' test results for substance i and concentration j . The standard deviation per substance i and concentration j is given by

$$\sigma_{i,j} = \sqrt{\frac{\sum_{l=1}^{r_{i,j}} (y_{i,j,l} - \bar{y}_{i,j})^2}{(r_{i,j}-1)}}. \quad (4.3)$$

Table 4.1: Notation for calculating the pooled standard deviation SD_p of experimental results per substance and concentration (dose in case of the LLNA) according to Eq. 4.2

| Notation | Explanation |
|-----------------|---|
| T | Classification threshold in a testing method's prediction model |
| i | Substance ($i = 1, \dots, n$) |
| n | Number of substances |
| j | Concentration tested per substance i ($j = 1, \dots, k_i$) |
| k_i | Number of concentrations per substance in the sample |
| $r_{i,j}$ | Number of runs per substance i and concentration j |
| l | Run per substance i and concentration j ($l = 1, \dots, r_{i,j}$) |
| $y_{i,j,l}$ | Test result of substance i , concentration j and run l |
| $\bar{y}_{i,j}$ | Arithmetic mean of test results for substance i and concentration j |

The BR in case of the DPRA was quantified using results from repeatedly testing $n = 42$ substances, yielding 446 runs (i.e. individual results) for different concentrations including the positive control (see Appendix A, Table A1), performed in a GLP-certified laboratory of BASF SE. The cysteine depletion of a given run was combined with the lysine depletion of a random run. This revealed pairs of cysteine and lysine depletion values. For each pair we determined the mean peptide depletion per substance and concentration. The BR was then calculated for test results revealing mean peptide depletion values between 3.38% and 9.38%. The BR in the prediction model of LuSens was calculated for test results from $n = 26$ substances, including the positive and negative control, yielding 2206 runs (i.e. individual results) from different concentrations (see Appendix A, Table A2). Again, experiments were conducted in a GLP-certified laboratory of BASF SE (using the Multimode Reader TriStar2 luminometer - Berthold Technologies, Germany), applying the classification threshold $FI = 1.50$. For each experiment, the BR was calculated for test results with luciferase fold-induction (FI) values up to 3.00 ($FI < 3.00$), and for test substance concentrations affording at least 70% relative viability. For assessing whether an unknown substance should be classified borderline (or not) we first investigated whether results from each concentration tested in a certain run fell into the BR (or not). Following to this we defined a decision rule for concluding on the overall assessment across runs within experiments (see Section 4.2.3, Table 4.2). In a second step we defined a decision rule was determined guiding conclusions on the overall assessment (borderline/non borderline) across experiments (Section 4.2.3, Table 4.3).

The BR around the classification threshold of the h-CLAT was calculated for test results from testing $n = 13$ substances during routine (in house) test applications, yielding 528 runs

(i.e. individual results) covering different concentrations (see Appendix A, Table A3). The BR was quantified for test results of fold inductions (FI) up to 3.00 fold for CD54 ($CD54FI < 3.00$) and up to 3.00 fold for CD86 ($CD86FI < 3.00$) for substance concentrations affording at least 50% relative viability. Since according to the testing protocol (OECD, 2016d) an experiment does not have to be conducted for different concentrations we first classified the result from each run (positive, negative, borderline). In a follow-up step we defined a decision rule to conclude on the overall assessment across experiments (see also Section 4.2.3).

Finally, the BR of the LLNA was quantified for test results from testing the $n = 22$ performance standard (PS) substances (ICCVAM, 2009) according to good laboratory practise (GLP), yielding 479 runs (i.e. individual results) for substances at different concentrations, applying the classification threshold of $SI = 3$ (see Appendix A, Table A4). For determining the BR only those chemicals with an SI in the range between ($2 \leq SI \leq 4$) were considered. The reason is that chemicals with an SI far above or below $SI = 4$ were observed to be of no or only marginal impact on the BR.

4.2.3 Decision rules for identifying borderline substances tested with individual non-animal methods

Given the BR around the classification threshold of each testing method we defined decision rules which guide, corresponding to the prediction model applied in each method, the identification of borderline substances in an experimental sample.

In case of the DPRA, substances for which the mean depletion rate was found to be between 4.86% and 7.90% were defined borderline. The prediction model of LuSens as described in Ramirez et al. (2014) requires that two consecutive concentrations per run reveal results above (below) the classification threshold in order to assess the test substance as positive (negative). Thus, a complete experiment reveals at least two independent results. If they are discordant, a third run has to be conducted and the conclusion on a substance's skin sensitisation potential is based on the majority outcome. For LuSens we established decision rules for determining the final result across all concentrations considered in repeated runs of an experiment (Table 4.3). Given the BR around the classification threshold of the LuSens prediction model ($1.26 \leq FI \leq 1.74$, see also Table 4.2) the outcome of an experiment was concluded to be positive (negative) if all results were above (below) the upper (lower) margin of the BR. If the first concentration (denoted x in Table 4.2) gave a negative result and the consecutive concentration ($x+1$) was either tested borderline or negative, it was concluded that the overall test outcome is negative. If LuSens revealed a

Borderline range

borderline result for a certain concentration x and the follow-up concentration $(x+1)$ was tested borderline or positive, the substance was decided to be a borderline substance.

Table 4.2: Decision rule for concluding on the overall test result of LuSens after two consecutive concentrations in a run

| Non-animal testing method results | Concentration x | Concentration ($x+1$) | Overall test result |
|--------------------------------------|-------------------|----------------------------|---------------------|
| | N | N | N |
| | P | P | P |
| | B | B | B |
| | N | B | N |
| | B | P | B |

N: Negative test result, indicating that a substance has not a keratinocyte activating potential;
P: Positive test results, indicating that a substance has a keratinocyte activating potential;
B: Substances which fall within the BR.

In case of the h-CLAT, at least one of the test results of either the CD54 expression or the CD86 expression from at least one of the runs in an experiment has to fall into the BR for qualifying an experimental result as borderline. Hence, the conclusion on the overall result of the experiment (positive, negative) is based on results from just one concentration.

Finally, we established a decision rule allowing to conclude on the overall test result across experiments. This was necessary because the testing protocols for LuSens and the h-CLAT require conducting two or more runs in order to classify a substance according to the results. The decision rules for the final conclusion on a substance's skin sensitisation potential across all possible runs conducted are shown in Table 4.3:

Table 4.3: Decision rules for LuSens and the h-CLAT to conclude on the overall test result after seeing results from repeated runs

| Number of runs ^a | 1 | 2 | 3 | 4 | Overall conclusion |
|----------------------------------|---|---|---|---|--------------------|
| Non-animal testing method result | N | N | N | N | N |
| | P | P | P | P | P |
| | N | N | B | B | B |
| | P | P | B | B | B |
| | N | N | P | B | B |
| | P | P | N | B | P |
| | N | P | B | B | B |
| | N | N | B | - | N |
| | P | P | B | - | P |
| | N | P | B | - | B |

N: Negative test result, i.e. a substance does not have a keratinocyte activating potential for LuSens or a dendritic cell activating potential for h-CLAT;

P: Positive test result, i.e. a substance has a keratinocyte activating potential for LuSens or a dendritic cell activating potential for h-CLAT;

B: Substances for which test results fall within the BR for either LuSens or h-CLAT.

^a Test results 1, 2,3 and 4 don't not imply fixed combinations.

4.2.4 Decision rules for identifying borderline substances tested with the “2 out of 3” ITS

Considering the BR of the prediction models of non-animal testing methods changes the possible outcomes of each method to be negative, positive, or borderline/ambiguous. Since test results of borderline substances can (by definition) not unambiguously be denoted positive or negative the respective substances cannot be compared with results from a reference animal test in order to conclude whether the test result is FP (i.e. erroneously classified as positive) or FN (i.e. erroneously classified as negative). The skin sensitisation potential is, however, assessed by a combination of the results of non-animal testing methods addressing different steps of the adverse outcome pathway (Jaworska, 2016; Kleinstreuer et al., 2016; Strickland et al., 2016). One of the simplest, yet successful, ways to do this, is the “2 out of 3” ITS (Bauch et al., 2012; Urbisch et al., 2015a). The “2 out of 3” ITS uses dichotomised results of individual non-animal testing methods (i.e. positive or negative). If a borderline/ambiguous outcome of an individual testing method is considered in the “2 out of 3” ITS, its overall conclusion of the skin sensitisation potential of a test substance may as well be borderline/ambiguous (or negative or positive). The “2 out of 3” ITS assigns equal weights to each testing method. Hence, the order of results of the individual methods does not matter. Consequently, one testing method yielding a borderline/ambiguous result will not change the overall result of the “2 out of 3” ITS, if the other two methods provided concordant – negative

or positive – results. If test results of prediction models of two non-animal testing methods fell into the BR, the overall outcome was borderline likewise, if the three methods yielded positive, negative and borderline/ambiguous results, respectively. Table 4 lists the overall outcome of the “2 out of 3” ITS depending on the results of the prediction models of the individual non-animal testing methods.

Table 4.4: Decision rules to conclude on the overall result using the “2 out of 3” ITS when considering borderline substances in individual non-animal testing methods

| Non-animal testing methods ^a | First test result | Second test result | Third test result | Overall conclusion |
|--|-------------------|--------------------|-------------------|--------------------|
| Non-animal testing method results¹ | N | N | N | N |
| | P | P | P | P |
| | B | B | B | B |
| | N | B | N | N |
| | P | B | P | P |
| | N | N | B | N |
| | P | P | B | P |
| | N | B | B | B |
| | P | B | B | B |
| | N | B | P | B |

N: Negative test result, i.e. a substance does not have a peptide reactivity potential for DPRA or a keratinocyte activating potential for LuSens or a dendritic cell activating potential for h-CLAT;

P: Positive test result, i.e. a substance has a peptide reactivity potential for DPRA or keratinocyte activating potential for LuSens or a dendritic cell activating potential for h-CLAT;

B: Substances which fall within the BR for either the DPRA, LuSens or the h-CLAT.

^a The order of test results does not imply the order of performing the non-animal testing methods.

4.3 Results

4.3.1. Quantification of the borderline range (BR) for the DPRA, LuSens, the h-CLAT and the LLNA

To quantify the BR around the classification threshold we used test results from substances tested with the non-animal testing methods DPRA, LuSens and h-CLAT, and from the LLNA, respectively. The number of substances with known skin sensitisation potential used to quantify the BR, the number of runs conducted per testing method, and the BR values of the testing methods’ prediction models, are shown in Table 4.5.

Table 4.5: Borderline range (BR) around the classification threshold of the animal test LLNA, and of the non-animal testing methods DPRA, LuSens, and h-CLAT

| Testing method | Number of substances (and runs) used for quantifying the BR* | Borderline range (BR) |
|----------------|---|-------------------------------------|
| LLNA | 22 (96 runs) | SI: 2.41 – 3.59 |
| DPRA | 42 (76 runs) | Mean peptide depletion: 4.86%– 7.9% |
| LuSens | 26 (473 runs) | Luciferase FI: 1.26–1.74 |
| h-CLAT | 13 (513 runs) | CD54 FI: 1.83–2.17 |
| | 13 (474 runs) | CD86 FI: 1.27– 1.73 |

BR: Borderline range;

SI: Stimulation index;

FI: Fold induction;

CD54, CD86: Cell surface markers;

See the Appendix A for a list of substances included in the experimental samples.

* For details about the composition of the samples see also Appendix A, Tables A1-A4.

If a substance is tested with any of the testing methods shown in Table 4.5, and if the result falls within BR of its prediction-model, a clear-cut conclusion about the substance's skin sensitisation potential is not possible. If, for instance, a substance tested with the DPRA reveals a mean peptide depletion between 4.86% and 7.90%, the result can neither be concluded to be negative nor to be positive. Instead, such test result would have to be qualified as “borderline” because results from repeated runs of the DPRA for this substance are likely to vary. For the specific case that test results are distributed normally, and that the BR is the mean of the distribution, the likelihood that a randomly selected substance is borderline is ~68%.

4.3.2 Identification of borderline substances in experimental samples tested with the non-animal testing methods DPRA, LuSens and h-CLAT, and with the animal test LLNA

Substances for which test results fell within the BR of the prediction models of the non-animal testing methods and the LLNA are listed in Table 4.6. We found that 6 out of 22 substances tested in the LLNA (i.e. 27%) were identified as borderline.

Of the borderline substances identified in the sample tested with the DPRA 9 revealed negative and 11 positive test results in the LLNA. Most substances with a negative test result were non-sensitisers based on LLNA potency classes. Four of the five substances for which test results were within the BR of LuSens revealed positive results in the LLNA. Of these, one was a weak, one a moderate and two strong sensitisers. Within the BR of h-CLAT all substances were positive when compared to the LLNA, three of which were weak sensitisers, one a moderate sensitiser, three were strong and one an extreme sensitiser.

Table 4.6: Borderline substances for the LLNA, the DPRA, LuSens, and the h-CLAT

| Testing method | Borderline substances | Sensitisation potential ^a in mice or humans (by conventional approach, assessed without <i>BR</i> ^b) | | Potency class (based on LLNA) |
|----------------|---|---|-------|-------------------------------|
| | | LLNA | Human | |
| LLNA | Salicylic acid ^c | N | N | Non-sensitiser |
| | Methyl salicylate ^c | N | N | Non-sensitiser |
| | Chlorobenzene ^c | N | - | Non-sensitiser |
| | Nickel chloride ^c | N | P | Non-sensitiser |
| | Phenyl benzoate ^c | P | P | Weak |
| | Methyl methacrylate ^c | P | P | Weak |
| DPRA | Salicylic acid ^c | N | N | Non-sensitiser |
| | α-Hexyl cinnamic aldehyde ^c | P | - | Weak / Moderate |
| | Geraniol | P | P | Non-sensitiser |
| | Benzyl alcohol | N | P | Non-sensitiser |
| | Tween 80 | N | N | Moderate |
| | 3-Dimethylamino propylamine | P | P | Weak |
| | Cis-6-Nonenal | P | - | Non-sensitiser |
| | Ethyl vanillin | N | - | Weak |
| | Undecylenic acid | P | P | Moderate |
| | 2-methoxy-4-methylphenol | P | - | Non-sensitiser |
| | Ethyl benzoylacetate | N | - | Moderate |
| | Dihydroeugenol | P | - | Weak |
| | N,N-Diethyl-m-toluanimide | N | - | Non-sensitiser |
| | Penicillin G | P | P | Weak |
| | d,l-Citronellol | P | N | Weak |
| | Pentachlorophenol | P | P | Weak |
| | p-tert-Butyl-α-ethyl hydrocinnamal (Lilial) | P | P | Weak |
| | 1-Bromobutane | N | - | Non-sensitiser |
| | Fumaric acid | N | N | Non-sensitiser |
| | Glucose | N | N | Non-sensitiser |
| LuSens | 1-Butanol | N | N | Non-sensitiser |
| | Benzoyl peroxide | P | P | Weak |
| | 4-Allylanisole | P | - | Extreme |
| | Methyldibromo glutaronitrile | P | P | Strong |
| | Imidazolidinyl urea | P | P | Strong |
| h-CLAT | 4-phenylenediamine ^c | P | P | Strong |
| | Phenyl benzoate ^c | P | P | Weak |
| | Ethylene diamine ^c | P | P | Moderate |
| | Aniline | P | P | Weak |
| | Farnesal | P | - | Weak |
| | Methyldibromo glutaronitrile ^c | P | P | Strong |
| | p-Benzoquinone | P | P | Extreme |
| | Propyl gallate ^c | P | P | Strong |

^a Prediction based on (Urbisch et al., 2015a), human data were extracted from (Basketter et al., 2014);

^b N=negative, P=positive.

^c Performance Standards (PS) substances of the OECD TG no. 429 (ICCVAM, 2009; OECD, 2010).

In case of the DPRA the percentage of substances falling into the BR was 10%, 6% in case of LuSens and 20% in case of the h-CLAT.

4.3.3 Identification of borderline substances in the experimental sample tested with the “2 out of 3” ITS

We found four substances out of 40 (10%) of the substances tested with the “2 out of 3” ITS to be borderline (Table 4.7), which is equal or less than the percentages revealed in case of the DPRA and the h-CLAT individually. All substances were positive in the LLNA. Of these, one is a weak one is a moderate and two substances are strong sensitisers according to the LLNA potency classes.

Table 4.7: Borderline substances in the experimental sample tested with the “2 out of 3” ITS

| Borderline substances | Sensitisation potential ^a in mice or humans | | Potency classes (based on LLNA) |
|-------------------------------------|--|-------|---------------------------------|
| | LLNA | Human | |
| Phenyl benzoate | P | P | Weak |
| Ethylene diamine | P | P | Moderate |
| Methyldibromo glutaronitrile | P | P | Strong |
| Propyl gallate | P | P | Strong |

^a Prediction based on (Urbisch et al., 2015a) human data were extracted from Basketter et al., (2014).

4.4 Discussion

4.4.1 Identification of borderline substances and implications of the BR for assessing substances' skin sensitisation potential

The BR defines the area around a testing method's classification threshold within which repeated testing will likely show discordant results. That is, within the BR a testing method is not precise due to its intra-assay variability. Given the BR, conclusions about a borderline substance's skin sensitisation potential are not possible. If a substance reveals test results falling within the BR, further testing is required to allow for a robust discrimination between a positive and a negative test outcome. The probability of an unknown substances to reveal a borderline result depends on the distribution of test results. For the specific case that test outcomes are normally distributed, and that the classification threshold is the mean of the distribution, the probability of seeing a borderline result is $p = 0.68$. This may differ for other types of distributions. Clearly, irrespective of the distribution of test outcomes the precision of a testing method is the higher (lower) the smaller (larger) the BR.

In this study we quantified the BR for prediction models of three non-animal testing methods as the pooled standard deviation around the testing method's classification threshold, the animal test LLNA, and the “2 out of 3” ITS. We find that 6 out of 22 (i.e. 27%) of the performance standard (PS) substances tested with the LLNA fall into its BR. This is slightly

higher than results obtained from the variability assessment in Hoffmann (2015), which may be explained by considering that Hoffmann (2015) determined the BR from EC3 values.

For the DPRA 20 out of 199 (10%) substances were identified as borderline, of which four were positive and seven negative in the LLNA. Applying the BR concept to LuSens required two steps to identify borderline substances (i.e. BR quantification within and across runs of experiments, see Tables 4.2 and 4.3). LuSens has a stringent prediction model (Ramirez et al., 2014; Ramirez et al., 2016). This may be a reason why LuSens revealed a relatively small percentage of borderline substances (6%, i.e. 5 out of 79). The application of the BR concept to the prediction model of the h-CLAT revealed 8 out of 40 substances (20%) being borderline. It should be noted that the prediction model of h-CLAT (Bauch et al., 2012) does not require concordant test results in consecutive concentrations of the same run to conclude on the substance's skin sensitisation potential. Furthermore, concordant test results with either cell surface markers CD54 expression or CD86 expression from at least two runs within the same experiment are required to conclude on a positive or negative test result (OECD, 2016d). Compared to the h-CLAT, the prediction model of LuSens is more elaborate because for each run two consecutive concentrations must be tested to determine the final result. Consequently, the prediction model of h-CLAT (OECD, 2016d) identifies a larger number of positive results (Sakaguchi et al., 2010), which may explain why all borderline substances in the experimental sample of h-CLAT were sensitisers.

4.4.2 Precision of non-animal testing methods compared to the LLNA

Taking the percentage of borderline substances in an experimental sample as a measure of a testing method's limited precision, we observe that this is considerably higher for the LLNA (27%) compared to the DPRA (10%), LuSens (6%) and the h-CLAT (20%). While this might be an indication for a larger imprecision of the LLNA compared to non-animal methods, the evidence provided in our study is not conclusive because experimental samples used differed across testing methods (24 PS substances in case of the LLNA, 199 substances for the DPRA, 79 substances for LuSens and 40 substances for the h-CLAT, respectively). Further research is required to examine the influence of sample size and composition on the quantification of the BR.

Of the borderline substances in the experimental sample of the LLNA two (i.e. phenyl benzoate, methyl methacrylate) are weak sensitisers, and four (i.e. salicylic acid, methyl salicylate, chlorobenzene, nickel chloride) are non-sensitisers (Table 4.6). Most substances identified as borderline in the LLNA are also discussed in Kolle et al. (2013). Our study also

identified phenyl benzoate as borderline, causing the percentage of substances falling in the BR of the LLNA to be slightly higher (27%) compared to Kolle et al. (2013) (23%). Note, however, that Kolle et al. (2013) determined the BR by calculating coefficients of variation based on individual animal data and did not use pooled animal data in the LLNA.

None of the substances identified as borderlines in the LLNA was borderlines in LuSens, one substance (salicylic acid) was also identified as borderline in the DPRA, and one substance (phenyl benzoate) was identified as borderline in the h-CLAT.

4.4.3 Precision of the “2 out of 3” ITS

Following the testing protocols for the DPRA (OECD, 2015a), LuSens (Ramirez et al., 2014; Ramirez et al., 2016) and the h-CLAT (OECD, 2016d) a single testing method cannot be used to predict skin sensitisation potential as a standalone method. The “2 out of 3” ITS has been suggested as a suitable approach for the overall assessment of the skin sensitisation potential based on the results of three individual testing methods (Urbisch et al., 2015a). Applying the BR concept to the “2 out of 3” ITS (Urbisch et al., 2015a) revealed four borderline substances in a set of 40 (10%), which is lower than that of the LLNA (27%). Our results, therefore, may indicate that the precision of the “2 out of 3” ITS is higher compared to the LLNA. Again, this result has to be treated with care because the experimental sample of the LLNA differed from that of the non-animal testing methods used in the “2 out of 3” ITS. Notwithstanding, the “2 out of 3” ITS reduces the influence of borderline substances on the overall conclusion about a substance’s skin sensitisation potential for all cases where two of the three methods provide concordant results. This, in turn, increases the overall precision of the “2 out of 3” ITS compared to the precision of the individual non-animal testing methods.

4.5 Conclusions

Technical and biological variability of non-animal testing methods used for assessing skin sensitisation potential, and the animal test LLNA, influence the precision of these methods. It is important to recognise that neither the animal test LLNA, often considered “the gold standard”, nor non-animal testing methods perfectly predict effects in humans (due to limited accuracy) and do not always yield clear-cut results (due to limited precision). A testing method’s precision constraint caused by intra-assay variability can be captured by quantifying a BR around the classification threshold of the method’s prediction model, which are used to transform continuous experimental data into a dichotomous result, being either “positive” (indicating an effect) or “negative” (indicating no effect). Test substances for which results fall within the BR of a testing method could be assessed as positive or negative upon

re-testing; thus the result of the test is ambiguous. Quite obviously, any conclusion drawn from experimental data is constrained by uncertainties and this is often neglected in reporting the results. The BR may offer a simple and pragmatic way to take into account that not every experimental data allows for a definite conclusion. A measure of precision, such as the BR, should therefore be reported with every study result. Furthermore, when using prediction models which dichotomise data there should always be three potential outcomes: positive, negative or borderline. While the paper focused on skin sensitisation as a proof-of-concept case, the BR approach is a generic method and can be applied to other endpoints, tests, and ITSs. Further research should, for example, quantify the BR for a broader set of (non-animal) testing methods, and should also address the impact of the size and composition of experimental samples on the BR. Moreover, examining the precision of testing methods for continuous endpoints deserves further attention in order to provide complementary insights into testing methods' precision regarding potency assessment (Slob, 2016).

Another important issue for further research and discussion is how to deal with borderline test results in a regulatory context. One possible option could be to define borderline results per default as positive results. However, this would imply that the upper part of the BR is factually ignored. Alternatively, one could require additional testing. Decision-theoretic approaches such the Bayesian Value-of-Information approach introduced in Leontaridou et al. (2016) can help to determine the optimal follow-up test in a systematic and transparent way. Finally, the question how borderline substances impact testing methods' predictive performance deserves further attention. Since for borderline substances the overall conclusion on their hazardous potential remains inconclusive, they cannot contribute unambiguously to the evaluation of a testing method's accuracy. Ignoring a substance's borderline result will, therefore, cause either over- or underestimation errors of, for example, a testing method's sensitivity or specificity. Exploring the size and direction of this impact for different non-animal testing methods, and analysing the influence of the size and composition of experimental samples, will provide complementary insights into the implications of intra-assay variability.

5 Uncertainties in measures of predictivity: The impact of precision, sample size and sample composition on the predictive accuracy of non-animal methods for skin sensitisation⁴

The ability of non-animal methods to predict the outcome of in vivo testing is expressed in terms of a test's predictivity (or predictive accuracy) by comparing the results of both tests obtained with a given number of substances. The predictive accuracy depends on the sample size (i.e. the number of substances which were tested to determine it), the composition of the sample, and the precision of both methods. Non-animal methods use prediction models to transform continuous read-outs of the test into dichotomous results by applying threshold values above and below which the test substance is assessed as positive or negative. Due to intra-test variability the precision of any testing method is limited. This results in a "borderline range" rather than a clear-cut classification threshold. When calculating the predictivity of non-animal methods it is usually not taken into account that test results of substances falling into the borderline range are inconclusive. This chapter explores the impact of intra-test variability on the predictivity of non-animal testing methods for assessing skin sensitisation potential. We quantify the impact a method's limited precision on the predictive accuracy of the DPRA assay, the ARE-Nrf2 luciferase method (covered by LuSens), the h-CLAT assay, and a combination of these methods into the "2 out of 3" integrated testing strategy. In addition, we examine impacts of intra-test variability on testing methods' predictivity caused by limited precision in combination with varying composition and size of experimental samples. Our results underline that discrete "positive/negative" outcomes are of limited informational value for evaluations of non-animal testing methods' predictivity. Instead, information on the variability, and the upper and lower limits of accuracy metrics should be provided to ensure transparent assessments and comparisons of testing methods' predictivity.

⁴ Chapter 5 is based on the manuscript in preparation: Leontaridou M., Gabbert S., Landsiedel R., (2017) Uncertainties in measures of predictivity: The impact of precision, sample size and sample composition on the predictive accuracy of non-animal methods for skin sensitisation.

5.1 Introduction

It has been widely acknowledged that binary “positive/negative” or “yes/no” outcomes have limited informational value regarding the “true” accuracy of non-animal testing methods, i.e. the degree of agreement between experimental results obtained and a corresponding (animal) reference test. Several studies pointed to possible biases in non-animal testing methods’ accuracy metrics due to inter- and intra-laboratory variability (Agnese et al., 1984; Margolin et al., 1984; Hothorn, 2002; Hothorn, 2003), which hampers a transparent comparison of non-animal testing methods predictivity with that of the animal test. Likewise, previous research revealed that animal test results can be biased due to technical and biological variability (Weil and Scala, 1971; Worth and Cronin, 2001b).

Recent research has paid specific attention to the intra-test variability of methods used for assessing skin sensitisation potential. Specifically, for the classification of substances’ hazardous potential both animal and non-animal testing methods apply prediction models using defined threshold values that dichotomise continuous experimental results into binary, i.e. positive and negative, outcomes (van der Schouw et al., 1995; Hoffmann and Hartung, 2005). Results from binary classifications are used to determine a testing method’s predictive accuracy compared to a reference test (e.g. the LLNA, (OECD, 2010)). Comparing experimental results obtained with a non-animal testing method with animal data allows quantifying the fractions of substances revealing true positive (TP), true negative (TN), false positive (FP), or false negative (FN) results (Krzanowski and Hand, 2009). Based on these fractions, a non-animal testing method’s predictive accuracy, e.g. sensitivity, specificity, and concordance (also called “accuracy”) can be determined. Predictive accuracy metrics specify a non-animal testing method’s ability to correctly classify an unknown substance compared to the reference animal test.

For the LLNA (Kolle et al., 2013; Hoffmann, 2015; Dumont et al., 2016) analysed the variability of classifications caused by a dichotomisation of continuous read-outs into discrete “positive/negative” data. In particular, Kolle et al. (2013) determined a range around the classification threshold within which the LLNA reveals discordant results in repeated applications. This range has been called “grey zone” (Dimitrov et al., 2016) or “borderline range” (BR) (Kolle et al., 2013). Hence, for substances yielding test results within the BR, clear-cut classifications of their skin sensitisation potential is not possible. This limits the LLNA’s precision, i.e. its ability to reveal concordant results in repeated applications. Leontaridou et al. (2017a) quantified the BR for the LLNA and the non-animal testing methods DPRA, LuSens and h-CLAT. Furthermore, their study determined borderline substances for

the “2 out of 3” integrated testing strategy (ITS), consisting of the non-animal testing methods mentioned above. The analysis showed that the number and percentage of substances considered borderline can be significant.

Clearly, substances with ambiguous hazard classification cannot contribute to determining a testing method’s predictive accuracy. As a consequence, ignoring the BR in a testing method’s prediction model – and, hence, the limited precision – may bias the assessment of classification accuracy. This hampers meaningful comparisons of accuracy metrics between non-animal testing methods and the reference animal test, e.g. for regulatory validation purposes. Besides the specification of the classification threshold, testing methods’ accuracy depends on the size and composition of experimental samples. Apart from using defined reference substances (denoted “proficiency chemicals”) (see Annex 2 of the DPRA OECD guideline (OECD, 2015a), Annex 2 of the ARE-Nrf2 luciferase method OECD guideline (OECD, 2015b), and Annex 2 of the h-CLAT OECD TG no. 442E (OECD, 2016d)), the composition and the size of the experimental samples depend on various considerations, e.g. the availability of robust reference data (i.e. *in vivo* and human data if applicable), the number of substances falling into each of the sensitisation potency classes and the number of sensitisers and non-sensitisers (ECVAM, 2012; ECVAM, 2013). Hence, the composition and the number of substances included in experimental samples can vary considerably. Furthermore, there is no defined minimum number of substances below which an experimental sample would be considered insufficient for robust evaluations of non-animal testing methods’ predictive accuracy. This induces additional bias, which can even interact with biases caused by testing methods’ limited precision.

So far, however, the impact of possible biases in the calculation of the abovementioned predictive accuracy metrics has not been systematically analysed. The aim of Chapter 5 is to fill this gap. We examine the impact of the limited precision on sensitivity, specificity and concordance of the non-animal testing methods DPRA, LuSens, and the h-CLAT. Currently, none of these methods is considered to provide sufficient information for classification as a standalone method (Mehling et al., 2012; Reisinger et al., 2015; ECHA, 2016). Combinations of these methods, for example the “2 out of 3” ITS (Bauch et al., 2012; Urbisch et al., 2015a), are assumed to provide sufficient information for concluding on a substance’s skin sensitisation potential. We therefore also include the “2 out of 3” ITS in the analysis.

The impact of classification bias is analysed in four steps: First, we examine the impact of non-animal testing methods’ limited precision on predictive accuracy metrics. This is done by comparing sensitivity, specificity and concordance derived from experimental samples

including borderline substances (i.e. the complete sample) with accuracy values obtained after borderline substances were excluded (i.e. the reduced sample). Second, we apply non-parametric bootstrapping (Wehrens et al., 2000) to create randomised experimental samples for every non-animal testing method considered. This generates distributions of sensitivity, specificity and concordance. Quantifying the mean, the standard deviation, and the 95% confidence interval for all accuracy metrics and compare the mean accuracy metrics to those from deterministic samples illustrate the impact of sample composition on classification bias. Third, we examine the joint impact of limited precision and sample composition by comparing accuracy metrics from randomised complete samples (i.e. including borderline substances) with those retrieved from reduced samples (i.e. excluding borderline substances). Finally, the joint impact from variations of sample composition, size and limited precision on classification accuracy is analysed.

The remainder of the chapter is structured as follows. Section 5.2 presents the methodological approach. We briefly explain the method to assess precision and document the experimental datasets used. In addition, we explain the scenarios for analysing the impact of classification bias on accuracy metrics. Section 5.3 discusses results from examining the impact of limited precision, sample size and sample composition on the predictive accuracy of non-animal testing methods. Section 5.4 discusses implication from our findings for assessing and comparing accuracy across non-animal testing methods, and between non-animal testing methods and the animal test. Section 5.5 concludes.

5.2 Materials and methods

5.2.1 Non-animal testing methods for assessing skin sensitisation potential

Skin sensitisation is a key endpoint for safety evaluations of new and existing substances in different regulatory frameworks of the European Union (e.g. the REACH legislation (EC, 2006), the Cosmetics regulation (EC, 2009)). Skin sensitisers cause allergic responses after contact (UNECE, 2011), from which about 15% to 20% of the population suffers at least once in a lifetime with increasing prevalence (Thyssen et al., 2007; Peiser et al., 2012). Responding to the urgent need to minimise animal testing, several non-animal testing methods and integrated testing strategies have been developed (Mehling et al., 2012; Reisinger et al., 2015; Urbisch et al., 2015a). Of these, the Direct Peptide Reactivity Assay (DPRA) (Gerberick et al., 2004; Gerberick et al., 2007), and the ARE-Nrf2 luciferase method covered by KeratinoSens™ (Natsch et al., 2011), were validated by European Centre for Validation of Alternative Methods (ECVAM), and OECD test guidelines 442C and 442D (OECD, 2015a; OECD, 2015b) have been adopted. The ARE-Nrf2 luciferase method is also covered by LuSens (Ramirez et al., 2014;

Ramirez et al., 2016), which is currently under validation by ECVAM. The human cell line activation test (h-CLAT) (Ashikaga et al., 2006; Sakaguchi et al., 2006; Ashikaga et al., 2010; Sakaguchi et al., 2010) has recently been validated by ECVAM and is described in OECD TG no. 442E (OECD, 2016d). The DPRA, KeratinoSens™ or LuSens and the h-CLAT cover the three “key events” of the skin sensitisation adverse outcome pathway (AOP) (OECD, 2012b; OECD, 2012c).

In case of the DPRA (Gerberick et al., 2004; Gerberick et al., 2007), depletions of two model peptides containing a cysteine- or lysine residue as a reactive nucleophilic centre are measured after incubation with a test substance. If the mean cysteine- and lysine- peptide depletion is above 6.38%, when compared to depletion in the reference control, the test result is positive and the substance is considered to be peptide reactive.

For determining the keratinocyte activating potential induced by LuSens (Ramirez et al., 2014; Ramirez et al., 2016), the luciferase induction after treatment with a test substance is assessed relative to concurrent vehicle controls. If a statistically significant fold induction (FI) of the luciferase activity is above 1.5, at relative cell viabilities of at least 70%, the test result is positive and the substance is considered to have a keratinocyte activating potential.

In case of the h-CLAT (Ashikaga et al., 2006; Sakaguchi et al., 2006; Ashikaga et al., 2010; Sakaguchi et al., 2010) the induction of the expression of the cell surface markers CD54 and CD86 is measured after treatment with a test substance, relative to concurrent vehicle controls in immortalized human monocytic leukemia THP-1 cells as a surrogate of DCs. If at least a two-fold induction of the CD54 expression and/or a 1.50-fold induction of CD86 expression are observed at relative cell viabilities of at least 50%, the test result is positive and the substance is considered to indicate a dendritic cell activating potential.

The “2 out of 3” ITS (Bauch et al., 2012; Urbisch et al., 2015a) combines test results from the DPRA, the ARE-Nrf2 luciferase method (covered by either LuSens or KeratinoSens™) and the h-CLAT. Equal weights are attached to each of the non-animal testing methods, which capture the three key events of the skin sensitisation AOP. The overall classification of a substance is determined by the majority of concordant test results from the DPRA, LuSens or KeratinoSens™ and the h-CLAT, respectively.

5.2.2 Quantification of the borderline range

Due to biological and technical variability we can identify a borderline range (BR) around the classification threshold within which test results can neither be classified positive or negative, but they must be reported as “non-conclusive” or “ambiguous” (Leontaridou et al., 2017a). The BR, therefore, constraints a testing method’s precision. Leontaridou et al. (2017a)

quantified the BR around a testing method's classification threshold (T) as the pooled standard deviation (SD_p) of test results from runs of testing methods, pooled across substance i and concentration j used. The quantification of the borderline rage is described in detail in Chapter 4 Section 4.3.1 of this thesis.

Using experimental results for the DPRA, LuSens, the h-CLAT and for the "2 out of 3" ITS published in (Bauch et al., 2012; Urbisch et al., 2015a), Leontaridou et al. (2017a) identified 20 substances out of 199 tested with the DPRA (10%), 5 out of 79 tested with LuSens (6%), and 8 out of 40 tested with the h-CLAT (20%) to be borderline when compared to results from the LLNA. For the "2 out of 3" ITS, 4 of 40 substances (10%) were identified as borderline.

5.2.3 Calculation of testing method's accuracy metrics

The predictive accuracy of non-animal testing methods and of the "2 out of 3" ITS approach was determined by means of three accuracy metrics, i.e. sensitivity, specificity, and concordance. Using standard 2x2 contingency tables (Cooper et al., 1979) the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) experimental test results of each non-animal testing method was determined when compared to LLNA and human data. The accuracy metrics sensitivity (Se), specificity (Sp) and concordance (Con) were quantified as follows:

$$Se [\%] = \frac{TP}{TP+FN} * 100, \quad (5.1)$$

$$Sp [\%] = \frac{TN}{TN+FP} * 100, \quad (5.2)$$

$$Con [\%] = \frac{TP+TN}{TP+TN+FP+FN} * 100. \quad (5.3)$$

5.2.4 Scenarios for analysing the impact of limited precision, sample size and sample composition on non-animal methods' predictive accuracy

To examine the impact of limited precision, sample size and sample composition on classification bias of non-animal testing methods we defined different scenarios, which are summarised in Table 5.1.

Table 5.1: Scenarios for assessing the impact of variations of precision, sample composition and sample size on non-animal methods' predictive accuracy metrics

| | Experimental sample sample size = n | Randomised sample sample size = n | Randomised sub-samples sample size < n |
|--|--|--|--|
| Complete sample (including borderline substances) | Scenario 1: Impact of limited precision on predictive accuracy metrics | Scenario 2a: Impact of varying sample composition on predictive accuracy metrics | Scenario 3a: Joint impact of varying sample composition and sample size on predictive accuracy metrics |
| Reduced samples (excluding borderline substances) | | Scenario 2b: Joint impact of varying sample composition and limited precision on predictive accuracy metrics | Scenario 3b: Joint impact of varying sample composition and size, and limited precision on predictive accuracy metrics |

First, we determined sensitivity (Eq. 5.1), specificity (Eq. 5.2) and concordance (Eq. 5.3) using the experimental datasets revealed for the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS, and compared to both the LLNA and to human reference data (Natsch et al., 2011; Bauch et al., 2012; Urbisch et al., 2015a) (Scenario 1 in Table 5.1). The composition of experimental samples is documented in Table B1-B4 in the Appendix B. The tables show experimental test results for the non-animal testing methods when compared to the LLNA and to human data.

Accuracy metrics were derived from experimental test results for the complete samples of substances (i.e. including the borderline substances) and for reduced samples (i.e. excluding the borderline substances). For ease of presentation, we confine the discussion on results revealed from experimental data compared to the LLNA as reference test. Results revealed from experimental samples using human data as reference are presented in the Appendix D. Table 5.2 shows the number of substances in the complete and the reduced samples used for calculating predictive accuracy metrics of the non-animal testing method and the “2 out of 3” ITS.

Table 5.2: Number of substances (n) in the experimental samples used for calculating predictive accuracy metrics of non-animal testing methods and the “2 out of 3” ITS *.

| | Complete samples (including borderline substances) | Reduced samples (excluding borderline substances) | Number and percentage of borderline substances |
|-----------------|---|--|---|
| DPRA | 199 | 179 | 20 (10%) |
| LuSens | 79 | 74 | 5 (6%) |
| h-CLAT | 40 | 32 | 8 (20%) |
| “2 out of 3”ITS | 40 | 36 | 4 (10%) |

* Experimental data compared to the LLNA as reference test.

Experimental data extracted from: Natsch et al., (2011); Bauch et al., (2012); Urbisch et al., (2015a).

Source: Leontaridou et al. (2017a).

Second, accuracy metrics were calculated for randomised samples. This captures uncertainty of test results due to varying sample composition. We determined accuracy metrics for the complete sample (i.e. including borderline substances, Scenario 2a in Table 5.1) and for the reduced sample (i.e. excluding borderline samples, Scenario 2b in Table 5.1). The latter offers a means to analyse the joint impact of sample composition and precision limitations on accuracy metrics. Randomisation was achieved by applying non-parametric standard bootstrap resampling analysis (Table 5.3). This method was used earlier by (Worth and Cronin, 2001b) to assess the variability of the Draize tissue scores. Our study applies a similar approach but focuses on the assessing the combined impact of varying sample composition and limited precision on non-animal methods' accuracy.

For every non-animal method and the "2 out of 3" ITS a set of $m = 10,000$ randomised samples (Efron and Tibshirani, 1993; Ostaszewski K. and Rempala G.A., 2000) was created by random replacement of the binary classifications obtained from experimental test results (Table 5.3, Step 1). The number of substances in randomised samples, denoted n , was equal to the number of substances in the complete and reduced experimental samples (column 2 and 3 in Table 5.2). Randomised samples were assumed to be independent and identically distributed (Wehrens et al., 2000). For all randomised samples we determined sensitivity, specificity and concordance according to Eq. 5.1-5.3. This revealed non-parametric distributions of sensitivity, specificity and concordance for the complete samples (i.e. including borderline substances), and for the reduced samples (excluding borderline substances, see also Table 5.3, Step 2 and 3). For every distribution we determined the mean and the standard deviation (SD) according to Eq. (5.4) and (5.5):

$$Mean_a = \frac{\sum_1^y a_y}{m}, \quad (5.4)$$

$$SD = \sqrt{\frac{\sum_1^y (a_y - Mean_a)^2}{m-1}}, \quad (5.5)$$

with a denoting the accuracy metric which is determined from the randomised sample (thus Se^* , Sp^* , Con^*), and y denoting the number of random samples ($m = 10,000$).

Table 5.3: Steps for conducting non-parametric standard bootstrap resampling analysis

| Step | Description |
|--------|---|
| Step 1 | Bootstrap resampling with random replacement of experimental test results from the individual non-animal testing methods and the "2 out of 3" ITS. |
| Step 2 | Quantification of a , thus sensitivity (Se^*), specificity (Sp^*) and concordance (Con^*) for the bootstrap sample. |
| Step3 | m -fold repetition of step 1 and 2; $m = 10,000$. |
| Step 4 | Calculation of the mean, the standard deviation and the 95% confidence interval of the distributions obtained for sensitivity (Se^*), specificity (Sp^*) and concordance (Con^*). |

In addition, we calculated confidence limits using the simple percentile method. Specifically, the 95% confidence interval (95% CI) was determined by the value corresponding to the 2.5% and 97.5% percentile in the bootstrap distribution of sensitivity, specificity and concordance, respectively (Table 5.3, Step 4).

Third, we assessed accuracy metrics for randomised sub-samples of varying sizes (Scenario 3a and 3b in Table 5.1) in order to analyse the combined impacts of uncertainty in sample size and composition, and of limited precision on non-animal testing methods' accuracy. Following the procedure outlined in Table 5.3 we calculated the mean, the SD and the 95% CI of the predictive accuracy metrics for each sub-sample including and excluding borderline substances. Sub-samples were $n = 10; 50; 100$; and 150 substances for the DPRA (with random replacement from the experimental sample consisting of 199 substances), $n = 10; 20; 40$ and 60 substances for LuSens (with random replacement from the experimental sample consisting of 79 substances), and of $n = 10$ and 20 substances for the h-CLAT and the "2 out of 3" ITS (with random replacement from the experimental samples consisting of 40 substances), respectively.

5.3 Results

5.3.1 Impact of precision uncertainty on accuracy metrics of the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS

In this section we present the results from analysing the impact of borderline substances on accuracy metrics of the DPRA, LuSens, the h-CLAT and the “2out of 3” ITS. Accuracy metrics were derived from experimental results using the LLNA as reference test. Results obtained from testing using human reference data are shown in Table D1-D3 in Appendix D.

Table 5.4: Impact of precision on predictive accuracy metrics of non-animal testing methods and the “2 out of 3” ITS*

| DPRA | | |
|---|-----------------|---------------|
| | | n =199 |
| Complete samples (including borderline substances) | Sensitivity [%] | 76 |
| | Specificity [%] | 72 |
| | Concordance [%] | 75 |
| | | n =179 |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 85 |
| | Specificity [%] | 80 |
| | Concordance [%] | 83 |
| LuSens | | |
| | | n =79 |
| Complete samples (including borderline substances) | Sensitivity [%] | 75 |
| | Specificity [%] | 70 |
| | Concordance [%] | 73 |
| | | n =74 |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 77 |
| | Specificity [%] | 69 |
| | Concordance [%] | 74 |
| h-CLAT | | |
| | | n =40 |
| Complete samples (including borderline substances) | Sensitivity [%] | 88 |
| | Specificity [%] | 87 |
| | Concordance [%] | 88 |
| | | n =32 |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 82 |
| | Specificity [%] | 87 |
| | Concordance [%] | 84 |
| “2 out of 3” ITS | | |
| | | n =40 |
| Complete samples (including borderline substances) | Sensitivity [%] | 85 |
| | Specificity [%] | 93 |
| | Concordance [%] | 88 |
| | | n =36 |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 82 |
| | Specificity [%] | 93 |
| | Concordance [%] | 86 |

* Experimental data compared to the LLNA as reference test.

Source: Own calculations.

5.3.2 Impact of uncertainty in sample composition and precision on accuracy metrics

As a result of the non-parametric bootstrapping procedure we received for each individual non-animal testing method a distribution of accuracy metrics. For the DPRA and LuSens we found predictive accuracy metrics to be normally distributed, while for the h-CLAT and the “2 out of 3” ITS we observed a left-skewed distribution of accuracy metrics (see Table C13-C24 in the Appendix C).

Randomisation causes the composition of substances in the samples to differ. Consequently, the number of borderline substances differed as well. The minimum and maximum number of substances in the reduced samples (i.e. after borderline substances were excluded) is shown in Table 5.5.

Table 5.5: Minimum and maximum number of substances (*n*) in randomised samples resulting from bootstrap resampling, after borderline substances were excluded*

| Randomised sample size (<i>n</i>) | DPRA | LuSens | h-CLAT | “2 out of 3” ITS |
|-------------------------------------|------|--------|--------|------------------|
| Min | 160 | 65 | 33 | 27 |
| Max | 194 | 79 | 40 | 40 |

* Experimental data compared to the LLNA as reference test.
Source: Own calculations.

Table 5.6 presents the mean and the SD (column 3), and the 95% CI (column 5) revealed from the distributions of sensitivity, specificity and concordance values. The table shows results obtained for the complete samples, i.e. samples including borderline samples (Scenario 2a in Table 5.1), reflecting the impact of variations in sample composition on accuracy metrics. Furthermore, the table documents the mean, the SD and the 95% confidence interval for distributions retrieved from the reduced samples (i.e. excluding borderline substances, scenario 2b in Table 5.1). This illustrates the joint impact of a varying sample composition and limited precision.

Table 5.6: Mean, Standard deviation (SD) and 95% confidence intervals of predictivity values of the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS determined from randomised experimental samples*

| | | Mean ± SD | 95%CI |
|---|-----------------|-----------|----------|
| DPRA | | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 76±4 | (69;83) |
| | Specificity [%] | 72±6 | (60;83) |
| | Concordance [%] | 75±3 | (69;81) |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 80±4 | (73;86) |
| | Specificity [%] | 74±6 | (62;86) |
| | Concordance [%] | 78±3 | (72;84) |
| LuSens | | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 75±6 | (63;87) |
| | Specificity [%] | 71±9 | (53;88) |
| | Concordance [%] | 73±5 | (64;83) |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 77±6 | (65;89) |
| | Specificity [%] | 69±9 | (51;87) |
| | Concordance [%] | 74±5 | (64;84) |
| h-CLAT | | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 88±7 | (74;100) |
| | Specificity [%] | 87±9 | (67;100) |
| | Concordance [%] | 87±5 | (78;98) |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 82±9 | (62;100) |
| | Specificity [%] | 87±9 | (67;100) |
| | Concordance [%] | 84±6 | (71;96) |
| “2 out of 3” ITS | | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 85±7 | (70;96) |
| | Specificity [%] | 93± 7 | (77;100) |
| | Concordance [%] | 87± 5 | (78;98) |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 82±8 | (64;96) |
| | Specificity [%] | 93±7 | (77;100) |
| | Concordance [%] | 86±6 | (74;97) |

* Experimental data compared to the LLNA as reference test.

Source: Own calculations.

5.3.3 Assessing the joint impact of uncertainty in sample size, sample composition and precision on accuracy metrics

Determining the mean, the SD and the 95% CI of accuracy metrics allows analysing the joint impact of sample size and composition on predictive accuracy metrics of the DPRA, LuSens, the h-CLAT and of the “2 out of 3” ITS (scenario 3a in Table 5.1). Results are shown in Table 5.7.

Table 5.7: Mean, SD and 95% CI of accuracy metrics revealed for the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS for complete samples (i.e. including borderline substances)*

| Sample size (<i>n</i>) | Sensitivity [%] | | Specificity [%] | | Concordance [%] | |
|--------------------------|-----------------|-----------|-----------------|------------|-----------------|-----------|
| | Mean ± SD | 95%CI | Mean ± SD | 95%CI | Mean ± SD | 95%CI |
| DPRA | | | | | | |
| 10 | 76±16 | (43; 100) | 71±29 | (0; 100) | 75±14 | (50; 100) |
| 50 | 76±7 | (61; 90) | 72±12 | (46; 93) | 75±6 | (62; 86) |
| 100 | 76±5 | (66; 86) | 72±8 | (55; 87) | 75±4 | (66; 83) |
| 150 | 76±4 | (68; 84) | 72±7 | (57; 84) | 75±4 | (68; 81) |
| 199 ^a | 76±4 | (69; 83) | 72±6 | (60; 83) | 75±3 | (69; 81) |
| LuSens | | | | | | |
| 10 | 75±17 | (38; 100) | 70±28 | (0.0; 100) | 73±14 | (40; 100) |
| 20 | 75±12 | (50; 100) | 70±19 | (33; 100) | 74±10 | (55; 90) |
| 40 | 75±9 | (57; 91) | 71±13 | (44; 92) | 73±7 | (60; 88) |
| 60 | 75±7 | (61; 88) | 70±10 | (50; 89) | 73±6 | (62; 85) |
| 79 ^a | 77±6 | (65; 89) | 69±9 | (51; 87) | 74±5 | (64; 84) |
| h-CLAT | | | | | | |
| 10 | 88±14 | (56; 100) | 87±19 | (40; 100) | 87±10 | (60; 100) |
| 20 | 88±9 | (67; 100) | 87±13 | (57; 100) | 87±7 | (70; 100) |
| 40 ^a | 88±7 | (74; 100) | 87±9 | (67; 100) | 87±5 | (64; 84) |
| “2 out of 3” ITS | | | | | | |
| 10 | 84±15 | (50; 100) | 93±15 | (50; 100) | 87±10 | (60; 100) |
| 20 | 85±10 | (63; 100) | 93±10 | (67; 100) | 88±7 | (70; 100) |
| 40 ^a | 85±7 | (70; 96) | 93±7 | (77; 100) | 87±5 | (78; 98) |

* Experimental data compared to the LLNA as reference test.

^a Number of substances in the experimental sample.

Source: Own calculations.

Finally, the parameters revealed for the distributions of sensitivity, specificity and concordance from reduced samples (i.e. after excluding borderline substances, scenario 3b in Table 5.1) offer insights into the joint impact of sample size variation and composition, *and* limited precision on predictive accuracy metrics (Table 5.8). Note that due to the randomisation of experimental samples the number of borderline substances within sub-samples could vary. Thus, similar to Table 5.5 we can determine the minimum and maximum number of substances for all subsamples after excluding borderline substances, which is shown in column 1 of Table 5.8. Distribution parameters of accuracy metrics (columns 2-4) capture the range of sample sizes per sub-sample.

Table 5.8: Mean, SD and 95% CI of accuracy metrics revealed for the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS for reduced samples (i.e. excluding borderline substances)*

| Max. and min. number of substances in randomised, reduced sub-samples | Sensitivity [%] | | Specificity [%] | | Concordance [%] | |
|---|-----------------|-----------|-----------------|------------|-----------------|-----------|
| | Mean ± SD | 95%CI | Mean ± SD | 95%CI | Mean ± SD | 95%CI |
| DPRA | | | | | | |
| 6-10 (<i>n</i> = 10) ^a | 80±16 | (43; 100) | 74±30 | (0; 100) | 78±14 | (50; 100) |
| 37-50 (<i>n</i> = 50) ^a | 80±7 | (65; 93) | 74±13 | (47; 95) | 78±6 | (66; 89) |
| 89-99 (<i>n</i> = 100) ^a | 80±5 | (70; 89) | 74±9 | (56; 90) | 78±4 | (69; 87) |
| 121-146 (<i>n</i> = 150) ^a | 80±4 | (72; 88) | 74±7 | (60; 88) | 78±4 | (71; 85) |
| 160-194 (<i>n</i> = 199) ^a | 80±4 | (73; 86) | 74±6 | (62; 86) | 78±3 | (72; 84) |
| LuSens | | | | | | |
| 4-10 (<i>k n</i> = 10) ^a | 77±18 | (40; 100) | 69±29 | (0.0; 100) | 75±14 | (44; 100) |
| 14-20 (<i>n</i> = 20) ^a | 77±12 | (50; 100) | 69±19 | (30; 100) | 74±10 | (53; 94) |
| 30-40 (<i>n</i> = 40) ^a | 77±8 | (59; 92) | 69±13 | (40; 93) | 74±7 | (60; 87) |
| 48-60 (<i>n</i> = 60) ^a | 77±7 | (63; 90) | 69±11 | (47; 89) | 74±6 | (63; 85) |
| 65-79 ^a (<i>n</i> = 79) ^a | 78±6 | (65; 89) | 69±9 | (50; 86) | 75±5 | (64; 84) |
| h-CLAT | | | | | | |
| 6-10 (<i>n</i> = 10) ^a | 82±20 | (33; 100) | 87±20 | (33; 100) | 84±13 | (56; 100) |
| 7-20 (<i>n</i> = 20) ^a | 82±13 | (50; 100) | 87±13 | (57; 100) | 84±9 | (65; 100) |
| 33-40 ^a (<i>n</i> = 40) ^a | 82±9 | (62; 100) | 87±9 | (67; 100) | 84±6 | (71; 96) |
| “2 out of 3” ITS | | | | | | |
| 6-10 (<i>n</i> = 10) ^a | 81±18 | (40; 100) | 93±15 | (50; 100) | 86±12 | (60; 100) |
| 12-20 (<i>n</i> = 20) ^a | 82±12 | (56; 100) | 93±10 | (67; 100) | 86±8 | (68; 100) |
| 27-40 ^a (<i>n</i> = 40) ^a | 82±8 | (64; 96) | 93±7 | (77; 100) | 86±6 | (74; 97) |

* Experimental data compared to the LLNA as reference test.

^a Number of substances in the complete sub-samples.

Source: Own calculations.

5.4 Discussion

5.4.1 Precision uncertainty

Acknowledging that borderline substances cannot be classified as “positive” or “negative”, we expected that predictive accuracy metrics derived from samples including borderline substances will differ from those where borderline substances were removed. Indeed, our results confirmed that accounting for the limited precision of the non-animal testing methods DPRA, LuSens, the h-CLAT, and the “2 out of 3” ITS, changed accuracy metrics (Table 5.4). However, this impact was not symmetric across all non-animal testing methods and the “2 out of 3” ITS. In particular, whereas for the DPRA, sensitivity, specificity and concordance increased after borderline substances were excluded from experimental samples, for the h-CLAT all accuracy metrics derived from test results compared to the LLNA decreased considerably but remained almost unchanged when derived from substances compared to human data (see Table D3 in the Appendix D). For LuSens we observed a small increase of sensitivity and concordance but even a slight decrease of specificity values, respectively. For the “2 out of 3” ITS we found a decrease of sensitivity, whereas specificity remained unchanged and concordance slightly decreased when assessed with substances compared to the LLNA. Similar results were observed for the accuracy metrics when test results of the non-animal testing methods were compared to human data (see Table D3 in the Appendix D).

For individual non-animal testing methods the size and direction of the impact on accuracy metrics, when considering limited precision, depends on the composition of experimental samples. In addition, it depends on whether test results for borderline substances are above or below the classification threshold. If, as in the case of the DPRA, more borderline substances revealed results below the classification threshold (thus they would be classified as “negative” when ignoring precision), excluding these substances increases the fraction of substances classified as “positive”, which in turn causes sensitivity to increase (see Eq. (5.3)). In contrast, test results of substances identified as borderline in the h-CLAT were all above the classification threshold (Leontaridou et al., 2017a). Hence, excluding these substances in order to correct for ambiguous classifications decreases specificity for the experimental sample with the LLNA as reference test (see Eq. (5.2)). In addition, since accounting for the BR changed the fractions of TP, TN, FP and FN classifications of the substances remaining in the sample, we also observed a slight decrease of sensitivity and concordance. In case of LuSens, only few substances were identified as borderline in the experimental sample (5 out of 79 substances, i.e. 6%; Leontaridou et al. 2017a). Due to the stringent prediction model of LuSens (Ramirez et al., 2014; Ramirez et al., 2016), the impact of

excluding these substances on the values of predictive accuracy metrics was only marginal. Likewise, the prediction model of the “2 out of 3” ITS, basing the overall conclusion about the skin sensitisation potential of a test substance on at least two concordant test results from the DPRA, LuSens or the h-CLAT, and assigning equal weights to each testing method, reduces the impact of borderline substances on predictive accuracy metrics.

5.4.2 Impact of uncertainty in sample composition and precision on non-animal methods’ predictivity

Determining accuracy metrics from randomised samples allows specifying their variation within the area under the probability distribution where predictive accuracy metrics are expected to lie. Depending on the composition of the randomised sample, accuracy metrics can be higher or lower compared to those quantified for the deterministic experimental samples. For all individual non-animal testing methods and the “2 out of 3” ITS the mean of the distributions of accuracy metrics corresponded to the values in the deterministic experimental samples.

Our results illustrate that accounting for limited precision in combination with randomised sampling increased the mean sensitivity, specificity and concordance of the DPRA, but did not affect the variation of accuracy metrics. For LuSens, mean sensitivity slightly increased, whereas mean specificity decreased, but the SD remained unchanged. For the h-CLAT and the “2 out of 3” ITS we observed a clear decrease of mean sensitivity and a slight decrease of mean concordance, respectively. Accounting for uncertainty of sample composition and precision also led to a higher SD for distributions of sensitivity and concordance. Hence, the overall uncertainty of these metrics increased. Furthermore, we found the 95% confidence interval to increase for all accuracy metrics and methods considered. This underlines that capturing the variation of sample composition and limited precision causes the variability of accuracy metrics to increase and, hence, different types of uncertainties underlying to non-animal testing methods’ accuracy can accumulate.

5.4.3 Impact of uncertainty in sample composition, sample size and precision on non-animal methods’ predictivity

Assessing the joint impact of varying sample size and composition, our results demonstrate that increasing the sample size decreases the variation of predictive accuracy metrics (given by the SD) and the 95% confidence limits for all individual non-animal methods and “2 out of 3” ITS. More specifically, for all individual methods and the “2 out of 3” ITS the SD of accuracy metrics from randomised sub-samples was found to be up to four times higher than the SD obtained from randomised full samples (i.e. including $n = 199$, 79, and 40

substances for the DPRA, LuSens and the h-CLAT respectively). Furthermore, the 95% CI of predictive accuracy metrics was considerably larger, indicating that for very small sample sizes a robust assessment of predictive accuracy metrics cannot be provided. Apparently, variations of sample size in combination with varying sample composition (Scenario 3a in Table 5.1) had only marginal impact on the mean values of accuracy metrics. Very similar results for the SD and the 95% CI were obtained when considering limited precision in addition to uncertainty of sample size and composition also (Scenario 3b in Table 5.1, results see Table 5.8). This implies that the impact of uncertainty in sample size and sample composition has a dominant impact on intra-test variability of accuracy metrics. However, including limited precision in the assessment changed the mean of accuracy metrics' distributions.

Finally for the DPRA, we observed a stabilization of the SD and the 95% CI values at samples sizes of $n \geq 100$ substances, irrespective of whether borderline substances were included or excluded from the samples. Our findings suggest, therefore, that predictive accuracy metrics of the DPRA are not sufficiently robust when derived from samples containing $n \leq 100$ substances. Our results may have implications for the interpretation of predictive accuracy metrics presented in other studies. For instance, experimental samples used for validating the DPRA and the h-CLAT included 21 and 24 substances, respectively. In both validation reports substances were tested three times with the DPRA (in three different labs) (ECVAM, 2012), and four times (in four different labs) with the h-CLAT (ECVAM, 2013). Although the assessment of a non-animal testing methods' predictive accuracy is only one component in a validation study, it is an important piece of information for concluding on a testing method's ability to provide correct classifications in relation to a reference test. Our findings may, therefore, stimulate a scientific and a policy debate about (the criteria for defining) minimum sample sizes.

5.5 Conclusions

Predictive accuracy metrics of all testing methods, i.e. non-animal methods and animal tests, suffer from different types of uncertainty, causing biased conclusions about substances properties. This chapter explored the impact of limited precision, variation of sample composition and variation of sample size on accuracy metrics of non-animal testing methods for skin sensitisation assessment. We analysed the impact for each individual type of uncertainty and for different combinations. The analysis was applied to experimental samples of substances tested with the DPRA, LuSens, and the h-CLAT. Furthermore, we included an integrated testing strategy composed by these individual methods, the "2 out of 3" ITS, in the

assessment. Considering limited precision (due to the BR around classification thresholds in non-animal testing methods' prediction models) changed accuracy metrics whenever experimental samples contain borderline substances.

Accounting for limited precision could have either a positive impact on accuracy metrics (i.e. accuracy metrics increase) or a negative impact (i.e. accuracy metrics decrease). Generally, we conclude that the direction and the size of this impact cannot be predicted from the outset, but depends on the composition of the sample of substances. In particular, the impact depends on the number of borderline substances on both sides of a method's classification threshold (i.e. the number of substances which would be classified as TP, TN, FP and FN if the BR is ignored) in relation to the number of substances which are not borderline.

Using randomised instead of deterministic experimental samples allowed determining probability distributions of accuracy metrics, illustrating their variation when the composition of samples is assumed to be uncertain. When considering the joint impact of varying sample composition and limited precision, the mean values of accuracy metrics, but also their SD (indicating the variation of data) and the 95% CI limits (indicating the spread of the variation) changed. Again, the size and direction of these changes was not pre-defined but depended on the fractions of substances on both sides of the prediction model's classification threshold. Finally, we found that the expected bias of accuracy metrics is highest if we account for uncertainty due to varying sample composition, sample size and limited precision. Thus, impacts of different types of uncertainty on non-animal testing methods' predictive accuracy accumulate.

Although the precise impacts differed across individual non-animal testing methods and the "2 out of 3" ITS, our results warrant a number of general conclusions. First, in order to avoid erroneous specifications of testing methods' accuracy metrics, the BR needs to be determined and substances with experimental results falling within this range should be detected. Furthermore, assessments of non-animal testing method's predictive accuracy should include complementary information about the potential over- and under-estimation error of accuracy metrics due to limited precision. This is particularly relevant because, as our results illustrate, there is a clear link between predictive accuracy metrics, limited precision and the composition and size of experimental samples. Determining the SD and 95% CI of the accuracy metrics is a useful way to report the uncertainties due to sample variation and to provide the area which encompasses the predictive accuracy metrics of testing methods if re-assessed using different experimental samples. Finally, for a more coherent assessment of the predictive accuracy of testing methods, experimental samples should be of a sufficient size

and of varying composition (i.e. with regard to TP, TN, FP and FN substances). This is particularly relevant for regulatory validation processes, which are decisive for the acceptance or rejection of non-animal testing methods. Determining the number of substances in an experimental sample which is considered sufficient is a matter of further research. Also, while this chapter focused on non-animal testing methods assessing skin sensitisation potential, it is important to explore the (combined) impacts of varying sample composition, sample size and limited precision on the predictive accuracy of testing methods for a broader set of endpoints. Such assessments must also include the reference animal tests to allow for comparisons of biases in testing methods' predictive accuracy between animal and non-animal testing methods. This is, in our view, a prerequisite for transparent and informative evaluations of testing methods.

6 Synthesis

This thesis introduced an economic perspective on the development of non-animal testing strategies. The research is embedded in the overall context of increasing information requirements for safety assessments of several tens of thousands of chemicals produced or marketed within the EU, the existing trade-offs between information requirements, constrained testing capacities and the policy objective to phase-out animal testing. From a conceptual and methodological perspective, the thesis offered a complementary contribution to toxicological research on developing non-animal testing strategies, and the evaluation of their performance compared to traditional animal tests. Moreover, the research presented in the thesis attempts to integrate fundamental toxicological concepts and approaches, e.g. for quantifying information outcomes derived from testing, into a comprehensive framework for efficient testing. Section 6.1 summarises the main findings from this research with regard to the research questions presented in the first chapter. Section 6.2 discusses the methodological approaches used, and a reflection of the general scientific and policy context to which this thesis contributes. Section 6.3 discusses limitations of the applied approaches and methods with regard to scope, underlying assumptions and data availability. Finally, section 6.4 concludes and Section 6.5 suggests topics for further research.

6.1 Answers to the research questions

RQ1: What are relevant criteria guiding the development of non-animal testing strategies for skin sensitisation potential and potency assessment?

The development of resource-efficient toxicity testing strategies has been driven by the need for providing fast, less costly and animal-free testing methods or strategies that offer adequate and sufficient information for hazard and risk assessment of chemicals. Focusing on the toxicological endpoint of skin sensitisation, Chapter 2 of this thesis reviewed the state-of-the-art regarding the development of non-animal testing strategies, and identified the criteria which guide the development of testing strategies as proposed in the recent toxicological scientific and policy literature.

Key findings revealed from this analysis can be summarised as follows: Throughout the past decade a large number of individual non-animal testing methods has been developed. Notwithstanding, there is general consensus between scientists and regulatory decision-makers that none of these methods can serve as a suitable replacement of the reference animal test (e.g. the LLNA or guinea pig based tests), if performed as standalone methods. For those non-animal methods which have been formally validated by the European Centre for

Validation of Alternative Methods (ECVAM; Italy) (e.g. the DPRA, the Are-Nrf2 luciferase method and the h-CLAT), testing protocols emphasise that the individual methods should be used in combination with supplementary information in order to provide suitable information for hazard identification and for potency assessment (ECHA, 2016). Therefore, the integration of information from different sources into battery or sequential combinations has been suggested as a promising solution for solving the problem of developing adequate and relevant information in an efficient way. In Chapter 2, we identified the criteria suggested in the toxicological literature for developing non-animal testing strategies. Furthermore, Chapter 2 introduced the economic perspective to the issue of “resource-efficient” testing and set the conceptual and informational criteria to optimise toxicity testing. Following to this, we provided a comprehensive qualitative evaluation on how these criteria were implemented into non-animal testing strategies for skin sensitisation potential and potency assessment. In Chapter 2, we analysed whether testing strategies suggested as “resource-efficient” in the literature can fulfil economic efficiency criteria.

The current state-of-the art regarding the development of non-animal toxicity testing strategies assessing skin sensitising properties of substances is characterised by the development of numerous toxicity testing strategies, aiming at integrating information from different non-animal testing methods for skin sensitisation. Both deterministic and probabilistic toxicity testing strategies are proposed for the assessment of skin sensitisation potential and potency. Schemes such as “Integrated Testing Strategies” (ITS), “Integrated Approaches to Testing and Assessment” (IATA), “Defined Approaches” (DA) and “Weight of Evidence” (WoE) approaches have been proposed, often in similar context, for the combination of information derived from different sources such as *in vitro* testing methods, read across or *in-silico* QSAR methods. A guiding rule, often proposed, for constructing testing strategies using different non-animal testing strategies is to follow the consecutive key events of the adverse outcome pathway (AOP). Following the key events of the AOP, supports the construction of testing strategies based on biological relevance criteria. In order to improve the economic efficiency of a testing strategy it is important to balance information gains with costs. We observed that the current toxicity testing strategies for the assessment of skin sensitisation aim at improving information gains without using animal tests, however, little attention has been given to the aspect of costs. Indeed, minimising testing costs and the “costs of making errors” is one of the criteria often mentioned in the efficiency criteria proposed in the literature. Our findings in Chapter 2 suggested, however, that testing costs are not

systematically incorporated into the design of non-animal testing strategies for skin sensitisation.

According to the criteria suggested in the toxicological literature, testing strategies should be coherent, transparent, hypothesis-driven, unambiguous and cost effective (Jaworska and Hoffmann, 2010; Rovida et al., 2015). However, seeing the development of “resource-efficient” testing strategies from an economic perspective, implies a set of conceptual criteria necessary to ensure optimisation in the development of toxicity testing. The conceptual criteria are: (i) the valuation of information gains and costs, (ii) the weighing mechanism for balancing information gains and costs from testing, (iii) the uncertainty assessment of both gains and costs and (iv) the stopping rule indicating when testing should stop. Furthermore, to ensure the appropriateness of the information gains and costs from performing testing methods as standalones or in a strategy, we set key informational criteria, i.e. predictivity, reliability and mechanistic understanding of information from testing methods and associated direct and indirect costs.

Next we investigated how these key conceptual and informational criteria are implemented on existing examples of testing strategies suggested as “Defined Approaches” in the latest OECD report (OECD, 2016c) for assessing skin sensitisation. We identified testing strategies using deterministic approaches such as “2 out of 3” ITS (Bauch et al., 2012; Urbisch et al., 2015a) and tiered testing strategies (van der Veen et al., 2014a) or probabilistic approaches such as Bayesian networks (Jaworska et al., 2010; Jaworska et al., 2013) and artificial neural networks (Hirota et al., 2013; Hirota et al., 2015). For this, economic approaches such as value of information (VOI) analysis, cost benefit analysis (CBA) or cost effectiveness analysis (CEA) can offer the tools for improving the economic efficiency in toxicity testing strategy development. Economic approaches offer guiding rules for developing testing strategies for the assessment of skin sensitisation. Balancing informational gains and expected economic losses from testing, reveals when an additional testing method should be performed and when testing should stop. The development of toxicity testing strategies should, therefore, not only be guided by principles from toxicology, but also by criteria, such as the costs of testing and the economic costs to society of making incorrect judgements.

RQ2: How can non-animal toxicity testing strategies for assessing skin sensitisation potential be optimised?

Chapter 3 provides a decision-theoretic framework for optimising sequential testing strategies. Using a Bayesian Value-of-Information (VOI) analysis approach allows for

balancing expected social welfare gains from testing with expected costs, including costs of making errors, for individual testing methods and any possible combination of methods into a testing strategy. A key assumption of the VOI approach applied is that testing has value if and only if it leads to welfare-improving decisions on chemicals' use. The VOI approach has several convenient features. First, it is a quantitative approach. The outcome of the VOI model is a testing method's or a testing strategy's expected value of test information (denoted EVTI), which is a monetary estimate of the expected social net benefit from testing. Comparing the EVTI across testing methods and testing strategies allows for ranking testing options according to their expected social net benefits. This offers the means for determining with which testing method to start a testing sequence, the number of testing methods that should be included in a testing strategy, and when to stop testing. Furthermore, it allows for comparing the EVTI of non-animal testing methods and strategies with that of an animal test. It also offers the opportunity to systematically examine if, and to what extent, non-animal testing methods for skin sensitisation should follow the order of key events in the skin sensitisation adverse outcome pathway (AOP), although this has repeatedly been suggested in the literature as a sufficient guiding rule for skin sensitisation testing. Second, the VOI model is a probabilistic approach. By applying Bayesian inference it allows for quantifying the uncertainty related to the outcomes of any test (including the "gold standard" animal test), and for assessing the remaining uncertainty of the outcomes from testing after new information (e.g. from a follow-up test) has become available. In addition, the Bayesian specification of the VOI model accounts for and allows updating a decision-maker's beliefs about the properties of a substance. Third, VOI analysis integrates, besides toxicological information, also relevant economic information such as testing costs, marketing gains from releasing a (safe) substance, forgone marketing benefits in case of an erroneous ban, and possible health damage costs (in our case direct and indirect costs arising from allergic contact dermatitis) from an erroneous release of a toxic substance.

The Bayesian VOI approach was applied to selected non-animal testing methods for skin sensitisation potential assessment (i.e. the DPRA, LuSens, KeratinoSens™, the h-CLAT, the OECD Toolbox and battery combinations of those methods), and the animal test LLNA. To explore the applicability of the model, we used the preservative Kathon CG as a proof-of-concept case. Though the empirical application was based on a number of simplifying assumptions (for example with regard to the assumed marketing volume of Kathon CG, the market price of the substance, or the decision-maker's risk attitudes, see also Section 6.3.2) the analysis offers a number of interesting and novel insights into the principles of developing

efficient sequential non-animal testing strategies for assessing the skin sensitisation potential of substances. First, we can conclude that the expected value of information from testing does not only depend on the information outcome from testing, but on the interplay of multiple parameters. These are a decision-maker's prior beliefs upon the hazardous properties of a substance, the predictive capacity of non-animal testing methods, the expected payoffs from marketing a hazardous or non-hazardous substance, and testing costs. Second, if a decision maker has strong beliefs that a substance is a potential skin sensitizer, or if a substance has a high sensitisation prevalence, or even a combination of strong beliefs that a substance is a sensitizer with a high skin sensitisation prevalence, increase the value of additional information from testing, because expected social costs of the release of a hazardous substance are high. We found that a 3-step sequential testing strategy consisting of the battery of the DPRA and LuSens, followed by the OECD Toolbox, and KeratinoSens™ as third testing method revealed the highest EVTI compared to all 236 sequential 2-test and 3-test testing strategies analysed. For low prior beliefs, (i.e. a decision-maker assumes that a substance is a non-sensitizer), we found the battery combination of the DPRA and LuSens to rank first. Third, our results underlined that both battery and sequential combinations of non-animal testing methods have a higher EVTI than the animal test, in this case the LLNA. One reason is that the predictive capacity of battery combinations of non-animal testing methods is usually higher than that of individual non-animal methods, which reduces the probability of adopting erroneous decisions. Another reason is that sequential testing strategies offer the possibility to save methods and, therefore, testing costs because a follow-up test will only be conducted conditional on the outcomes of testing methods at earlier stages of the testing strategy. Finally, we observe that the order of non-animal testing methods in a strategy does not have to follow the order of key events in the skin sensitisation AOP. Thus, covering all key events in the AOP is neither a necessary nor a sufficient condition for efficient testing. In contrast, depending on a decision-maker's prior beliefs it may be preferred to generate more information for the same key event in two consecutive steps of the sequential testing strategy.

RQ3: How do technical and biological variability of non-animal testing methods influence the precision of non-animal testing methods for assessing skin sensitisation potential?

In Chapter 4 we assessed the impact of biological and technical variability on the precision of non-animal testing methods assessing skin sensitisation potential. Given that any test – irrespective of whether it is an animal test, an *in vitro* method (using cell-cultures), or an *in chemico* method or an *in silico* method (using computational methods), is a simplified model

representation of the processes that are expected to happen in the human body, biological and technical variability are inherent characteristics of testing methods. For hazard classification, continuous experimental data resulting from testing are usually dichotomised into binary “positive/negative” or “hazardous/non-hazardous” data by means of applying a pre-defined classification threshold. Earlier studies, i.e. (Kolle et al., 2013; Dimitrov et al., 2016), has shown that biological and technical variability influence the ability of the animal test LLNA to provide clear-cut conclusions about a substance’s skin sensitising potential when test results fall close to the pre-defined classification thresholds. This has been expressed as the range on both sides of the classification threshold of the LLNA in which test results can be discordant. Thus if a substance tested with the LLNA reveals results which fall within this range, called “grey zone” or “borderline range”, drawing conclusions about skin sensitisation potential is not possible because repeated testing may ambiguously reveal either positive or negative results. The precision of non-animal testing methods is, similarly to the LLNA, influenced by the biological and technical variability. Therefore non-animal testing methods also have a borderline range around their classification threshold in which their ability to provide clear-cut conclusions on the skin sensitising properties of substances is not possible with sufficient confidence. Substances yielding test outcomes within the borderline range of testing methods may require further testing in order to avoid misclassifications. Quantifying the borderline range in the prediction models of non-animal testing methods is, therefore, a practical way to account for borderline test results and to unravel a testing method’s limited precision.

We quantified the borderline range for selected non-animal testing methods for assessing skin sensitisation, i.e. the DPRA, LuSens, and the h-CLAT, and for the “2 out of 3” ITS. The latter has been introduced as an integrated testing strategy consisting of three non-animal testing methods. Furthermore, we identified the substances in the experimental samples of these methods for which test results fell within the borderline range. Since each prediction model applied in each of the non-animal testing methods used is different, specific decision rules were defined to guide the identification of borderline substances. Our analyses revealed the following results: First, biological and technical variability do not only impact the prediction revealed from the animal test LLNA, but also predictions used for the non-animal testing methods considered in Chapter 4 for which we quantified their borderline range. The borderline range shows the area around the classification threshold, in which non-animal testing methods are not precise, i.e. they are likely to reveal discordant results in repeated applications. Second, the percentage of borderline substances in the experimental samples used in the DPRA, LuSens, the h-CLAT, as well as the “2 out of 3” ITS was less than that of the

reference animal test (LLNA). However, since the size and composition of experimental samples differed across non-animal testing methods, and between non-animal testing methods and the LLNA, our findings do not allow for comparisons between these methods (see also Section 6.4). For the “2 out of 3” ITS, we found that the percentage of substances yielding borderline test results was equal to the DPRA and lower compared to the h-CLAT. The reason is that the “2 out of 3” ITS applies a simple majority rule for concluding on a substance’s skin sensitisation potential, i.e. the classification is based on at least two concordant results. This majority rule ignores borderline substances in the sample of each testing method used in the “2 out of 3” ITS which is not considered for the classification decision. The majority rule of the “2 out of 3” ITS, therefore, allows excluding a borderline test result if the results from the other two testing methods fall outside the borderline range. Hence, one can conclude on the skin sensitisation potential of a substance even if individual methods revealed a borderline result.

RQ4: How do limited precision, sample size and sample composition impact the predictive accuracy of non-animal testing methods for skin sensitisation?

In Chapter 5 we analysed the uncertainties in predictive accuracy metrics of non-animal testing methods. We analysed the impact of the limited precision of non-animal testing methods on their predictive accuracy, i.e. a testing method’s ability to correctly detect an adverse effect in comparison to the reference animal test. Further, we analysed the impact of variations of sample size and sample composition on testing methods’ predictive accuracy. We examined these impacts both separately and in combination with limited precision. Common predictive accuracy metrics e.g. sensitivity, specificity and concordance (also called accuracy) are calculated by comparing binary “hazardous/non-hazardous” test results from non-animal testing methods (e.g. the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS) to those from a reference test (e.g. test results from the LLNA data, and human data when available) for a set of substances. As elaborated in Chapter 4, for substances yielding test results within the “borderline range” around the classification threshold of a testing method’s prediction model such clear-cut classifications into “hazardous/non-hazardous”, is not possible. Consequently, borderline substances cannot contribute to the assessment of a testing method’s predictive accuracy.

The impact of considering borderline substances on testing method’s predictive accuracy was analysed by comparing sensitivity, specificity and concordance quantified from experimental samples including borderline substances with values obtained when borderline

substances were excluded. Here we compared results obtained from experimental samples of a pre-defined sample size and composition with those obtained from randomised samples of a given size and of varying samples. Randomised samples were created using non-parametric bootstrap resampling analysis. The analysis was performed using the DPRA, LuSens, the h-CLAT and the “2 out of 3” ITS combining these three methods.

We can conclude that the limited precision of testing methods, reflected by the number of borderline substances in an experimental sample, can affect the predictive accuracy metrics. However, the magnitude and the direction of the impact on the predictive accuracy vary across non-animal testing methods. Generally, the impact depends on the number of borderline substances on both sides of a method’s classification threshold (i.e. the number of substances which would be classified as TP, TN, FP and FN if the *BR* is ignored) in relation to the number of substances which are not borderline. Whereas for the DPRA sensitivity, specificity and concordance increased after borderline substances were excluded from experimental samples, the accuracy metrics for the h-CLAT were considerably decreased after excluding the borderline substances from the experimental samples, when compared to the LLNA data. For LuSens we observed a small increase of sensitivity, but no change on concordance and even a slight decrease of specificity values. For the “2 out of 3” ITS, we found a decrease in sensitivity, whereas specificity remained unchanged and concordance even slightly decreased.

Determining predictive accuracy metrics from randomised samples revealed for each non-animal testing method considered distributions of sensitivity, specificity and concordance. Based on these distributions we quantified the 95% confidence intervals, the mean and the standard deviation of sensitivity, specificity and concordance, respectively. Mean values of accuracy metrics derived from experimental samples differed only marginally from mean accuracy metrics obtained from randomised samples. However, there can be considerable variation of accuracy metrics around the mean of the distributions. For the three non-animal testing methods the DPRA, LuSens and the h-CLAT, and for the “2 out of 3” ITS, this illustrates that the impact of considering borderline substances on the assessment of a non-animal testing method’s predictive accuracy can be either higher or lower based on the sample composition. Finally, we observed that the variation of accuracy metrics decreases with increasing number of substances in the samples used to assess the predictivity of testing methods (sample size). This underlines the relevance of using experimental samples of a sufficient size and of a balanced composition (i.e. including a balanced fraction of hazardous and non-hazardous substances) for ensuring a coherent assessment of non-animal methods’

predictive accuracy. Moreover, a transparent and coherent assessment of non-animal testing methods' predictive accuracy requires to identify and document borderline substances in experimental samples. Finally, due to appropriate standard deviations or confidence limits, indicating ranges of expected accuracy metrics, and not as point estimates.

6.2 General discussion

This section discusses the wider context of the research conducted in this thesis, briefly reflects on the methodological and modelling approaches used in the Chapters 2 to 5 and the policy relevance of the thesis.

6.2.1 Wider context of the thesis

In the overall context of chemicals' safety assessment and the development of efficient toxicity testing strategies scientific efforts have focused on developing new testing methods and conceptual approaches that facilitate the integration of information from different experimental and computational sources. The purpose is to support efficient hazard and risk assessment of the large numbers of chemicals produced worldwide, while reducing or avoiding the use of animal testing. The development and combination of non-animal methods, e.g. cell based methods (*in vitro*) and computational (*in silico*) approaches, has been considered a powerful approach for generating sufficient and relevant hazard information using less resources (i.e. time, testing costs, laboratory animals) than with traditional animal tests. In addition, integrating information from *in vitro* and *in silico* methods is considered to better cover the different key events in the AOP of certain endpoints compared to animal tests, provided that these events are known (Vinken, 2013). Still, a number of fundamental conceptual question have not been addressed systematically as yet, in particular how to optimally combine different information outcomes from *in vitro* and *in silico* methods, and how to arrive at conclusions on the hazardous properties of a substance.

Regarding the current status of developing integrated approaches to toxicity testing for skin sensitisation (Chapter 2) we identified different terminologies describing the effort to combine non-animal testing methods into strategies, for example "Integrated Testing Strategies" (ITS), "Sequential Testing Strategies" (STS), "Integrated Approaches to Testing and Assessment" (IATA), "Defined Approaches" (DA), or "Weight of Evidence approaches" (WoE) (Tollefsen et al., 2014; Rovida et al., 2015; Jaworska, 2016; Sauer et al., 2016). The recent reports from OECD, provide clear definitions and delineate the differences between the concept of IATA and any approach (i.e. DA) to integrate information from different sources using a fixed data interpretation procedure (OECD, 2016b; OECD, 2016c). Moreover, we

concluded that existing approaches focus predominantly on maximising information outcomes from testing. As we argue in this thesis, a prerequisite for optimal testing is to balance information gains from testing with costs for generating this information. This holds for both individual testing methods and for DAs to integrate information (testing strategies). Thus, the development of efficient testing strategies has clear characteristics of an economic optimisation problem (Yokota et al., 2004; Yokota and Thompson, 2004; Gabbert and van Ierland, 2010; Gabbert and Weikard, 2010; Gabbert and Weikard, 2013). While this thesis focuses on skin sensitisation, the problem of developing efficient non-animal testing strategies is generic and it applies to any health and environmental endpoint, e.g. liver toxicity (Daston et al., 2015; Gocht et al., 2015), aquatic toxicity (Villeneuve et al., 2014; Groh et al., 2015), fish toxicity (Nendza et al., 2014) or local tolerance endpoints such as skin corrosion and eye irritation (Sauer et al., 2016). We observed that several non-animal testing strategies have been developed based on purely toxicological criteria. Although theoretical approaches to the optimisation of testing strategies have been suggested earlier (Gabbert and van Ierland, 2010; Gabbert and Weikard, 2013; Norlén et al., 2014; Leontaridou et al., 2016), empirical applications – requiring in particular a quantification social gains and losses related to testing – have been largely lacking. The Bayesian VOI model suggested in Chapter 3 of the thesis is therefore a step forward to filling this gap. While the findings from our analysis refer to a specific human health endpoint (skin sensitisation), the insights into the conceptual features of sequential testing are generic. The thesis, therefore, lays the conceptual grounds for optimising toxicity testing for other toxicological endpoints, and also for more types of chemicals (e.g. nanoparticles).

To understand the implications of translating experimental readouts from testing into a final conclusion about the hazardous properties of a substance we need to carefully investigate the quality of test information. Based on our analysis on the precision of non-animal testing methods (Chapter 4) we observed that biological and technical variability can limit the precision of testing methods. This may lead to erroneous hazard classifications and to over- or underestimation errors of a testing method's predictive capacity. Again, using skin sensitisation as an illustrative case we showed that ambiguity of hazard classification is a general problem that applies to both animal tests (Kolle et al., 2013; Hoffmann, 2015; Dimitrov et al., 2016) and non-animal testing methods. Defining a quantifiable measure for this ambiguity, called "borderline range", Chapter 4 allows assessing its impact on the predictive accuracy of (non-animal) testing methods, i.e. the ability of a testing method to

provide test results concordant to those of a reference animal test for tested substances as described in test guidelines for non-animal testing methods e.g. TG no 442C (OECD, 2015a).

Currently, a testing method's predictive accuracy is determined by comparing dichotomised test results revealed from testing a pre-selected sample of substances to those obtained from a reference animal test. As we showed in this thesis, the borderline range and the number of chemicals which are likely to be misclassified due to the borderline range depend on the size and composition of these experimental samples. This raises the normative question of how to compose experimental samples in order to reduce ambiguity. Furthermore, it highlights the need to document complementary information about the uncertainty of test information, for example as part of validation reports of new testing methods. This holds in particular because the predictive accuracy of non-animal testing methods is compared with that of reference animal tests. The concept of using the borderline range as an additional measure to evaluate information from testing strengthens the appropriateness and trustworthiness of information based on which the use of substances is decided. As a consequence, this has implications for the quality of predictive accuracy metrics, also used as input into the optimisation framework developed in Chapter 3, and on the trustworthiness of hazard classifications for substances yielding test results close to the classification thresholds.

6.2.2 Methodological and modelling approaches

Although it is common sense among scientists and policy makers that testing costs should be minimised, the qualitative review conducted in Chapter 2 revealed that the development of existing testing strategies, and in particular of testing strategies for the assessment of skin sensitisation, has not been guided by economic efficiency criteria such as a mechanism for balancing gains and costs of testing. It was therefore concluded that there is a need for developing an optimisation framework to testing. Responding to this need we developed a decision theoretic framework using Bayesian VOI analysis that guides the process of combining different non-animal testing methods into sequential testing strategies in an efficient way. While the Bayesian VOI approach is not new per se (Yokota and Thompson, 2004), it offers a coherent, theory-based framework for determining (i) with which test to start in a testing strategy, (ii) how many non-animal testing methods to include, and (iii) when to stop testing. In our analysis the expected value of test information of non-animal testing methods and their combinations, EVTI, is compared to that of the reference animal test that has been considered a "first choice" in the traditional toxicity testing approach. The Bayesian VOI model which we suggested incorporates estimates of social benefits and losses from

decisions upon the use of a substance under either outcome from testing. Besides testing costs from performing a testing method, it captures monetised health damage costs due to skin sensitisation. Furthermore, it includes estimates of (foregone) marketing benefits. The gains and losses under either decision option are added in order to determine expected net benefits for every decision option (ban or release). We used Kathon CG as a proof of concept case for calibrating the model. The Bayesian VOI model offers a tool to link the predictive accuracy of testing methods with the expected payoffs from marketing substances by weighing the expected payoffs, with the probability of a substance to be hazardous/non-hazardous. This model can also be applied for a broader set of substances or endpoints.

The Bayesian VOI uses measures of a test's predictive accuracy (i.e. sensitivity and specificity) for characterising test information. These are a test's sensitivity (i.e. the proportion of skin sensitisers which were correctly classified) and specificity (i.e. the proportion of non-sensitisers which were correctly classified). Predictive accuracy metrics were derived by transforming continuous experimental readouts from testing into binary hazardous/non-hazardous outcomes, and by determining percentages of correctly and erroneously classified substances in the experimental sample as compared to a reference test. In a regulatory context, dichotomised test results are sufficient for hazard identification purposes required for any toxicological endpoint (EC, 2008; EC, 2016). The classification of substances as "hazardous/non-hazardous" is based on pre-defined thresholds applied to the continuous experimental readouts. However, using clear-cut thresholds may lead to an over- or underestimation of a testing method's sensitivity and specificity due to the biological and technical variability of the testing method. The biological and technical variability constraint the testing method's precision, i.e. its ability to reveal concordant results in repeated runs of the testing method. We determined the area around the threshold of testing methods in which test results are expected to be ambiguous. This area is called borderline range and we proposed this concept as an additional measure to evaluate the appropriateness of information from testing. To quantify the borderline range, we used substances that have been routinely tested with "in-house" experiments of testing methods. Routinely tested substances, in contrast to highly standardised and well-characterised substances used in validation experiments of the testing methods, are substances that are intended to be released into the market. Therefore, using substances routinely tested is a practical way to examine the precision of testing methods and reflect the precision of these methods in "in house" laboratory practices.

While the uncertainty underlying to predictive accuracy metrics of testing methods has been discussed earlier (Worth and Cronin, 2001a), the impact of this uncertainty on (i) the correct classification of hazardous and non-hazardous substances, and (ii) the conclusions on the predictive accuracy of testing methods, in particular non-animal testing methods, have not been systematically addressed. Our analysis, therefore, contributed to a more comprehensive understanding of the informational “value” of testing methods (Worth and Cronin, 2001a). As a first step we analysed the impact of the borderline range on non-animal testing methods’ accuracy metrics (Chapter 5). This was done by comparing accuracy metrics derived from test results of a sample of substances which include borderline substances, with those obtained after we excluded substances yielding borderline results. Excluding the substances for which test results fall within the borderline range implies that the concept of precision is considered before the decision makers translate experimental readouts into final conclusions on the classification of substances. Furthermore, we examined the impact of the borderline range for randomised substances samples, using non-parametric standard bootstrap analysis. Randomised samples of the same sample size as the experimental samples and of different sample sizes smaller than the size of experimental samples were generated. The bootstrap analysis allows defining 95% confidence intervals for the accuracy metrics of no-animal testing methods for both cases at which precision of a testing method is first considered and second ignored. Providing confidence intervals for accuracy metrics based on randomised samples means that decision makers can approximate the range of expected accuracy metrics (Worth and Cronin, 2001a). This provided insight into the impact of the borderline range on testing methods’ predictive accuracy when considered in combination with varying size and composition of experimental samples. The experimental samples used for evaluating a test’s predictive accuracy are usually composed on the basis of expert judgment or simply data availability, and thus they are neither randomised nor of a pre-defined size. Our analysis shows that documenting accuracy metrics as ranges rather than point estimates may be a more appropriate way of characterising test information uncertainty raised from number and composition of substances in the experimental samples. Predictive accuracy assessment of testing methods, irrespectively of the endpoint addressed, has been traditionally based on the comparison of test results derived from a testing method to those from a reference animal test. Therefore, the methodology used in Chapter 5 contributes to the current practices of assessing the predictivity of testing methods by suggesting to consider (i) the impact of limited precision on the classification of substances and (ii) the number and composition of substances used for assessing the predictivity of testing methods.

6.2.3 Policy relevance

During the past decade scientists and decision makers in industry or regulatory agencies have increasingly paid attention to the fundamental problem how toxicity testing strategies can be optimised. Regulatory frameworks such as the European REACH legislation (EC, 2006) and the Cosmetics Regulation (EC, 2009), but also policy reports such as the report of the U.S. National Academy of Sciences “Toxicity testing in the 21st Century: A Vision and a Strategy” (Krewski et al., 2010) have spelled-out detailed information requirements in order to ensure that the risks of toxic chemicals can be optimally controlled. Information requirements for hazard and risk assessment of large numbers of chemicals in a regulatory context have triggered the development of alternative testing methods. The development of testing methods has been focusing on creating robust, fast and cheap approaches aiming at reducing and eventually replacing the animal tests. This thesis contributes to this process by (i) surveying the state of the art and the conceptual criteria of developing non-animal testing strategies for skin sensitisation hazard and potency assessment, (ii) developing an economic framework to optimise non-animal strategies for skin sensitisation potential assessment, and (iii) by exploring uncertainties underlying to information outcomes of non-animal testing methods. The optimisation framework to developing non-animal testing strategies, using the Bayesian VOI analysis, offers a probabilistic tool to analyse the interplay between a decision-maker’s prior beliefs about the properties of a substance, expected information gains from testing, and costs. This acknowledges that the construction of optimised non-animal testing strategies must balance different and possibly competing objectives. Furthermore, it underlines that solutions to the general problem of “how to test” are not independent of the possible set of (regulatory) decisions upon the use of chemicals.

Although there has been an agreement among scientists and policy makers that animal testing should be phased out (Basketter et al., 2013; Jaworska, 2016; Sullivan, 2016; Worth and Patlewicz, 2016), there is still a controversial discussion on how this can best be achieved (Hartung et al., 2013; Rovida et al., 2015). Non-animal testing methods can be used in combination with other complementary information to support final conclusions on the skin sensitising properties of substances (ECHA, 2016), however, information from standalone non-animal testing methods may be sufficient for the classification of a substance into UN GHS category 1 (thus binary classification of a substances as “sensitiser/non-sensitiser”) under specific regulatory frameworks as explained in OECD test guidelines e.g. for the DPRA, the ARE-Nrf2 Luciferase method and the h-CLAT (OECD, 2015a; OECD, 2015b; OECD, 2016d). Scientific discussions, however, suggest integration of information from non-animal testing

methods for trustworthy classification of substances. As our results show the expected value of animal test information is lower than that of standalone non-animal testing methods. Performing non-animal testing methods combined in battery or sequential combinations increased the value of information of these methods even further. Moreover, our results underline that sequential testing strategies do not necessarily need to follow the order of key events of an AOP. Chapter 3, therefore, contributes to evaluating the relevance of non-animal testing methods for chemicals' hazard assessment.

In addition, results presented in this thesis contribute to the policy debate regarding the regulatory acceptance of non-animal testing methods. As we demonstrated in Chapters 4 and 5, using clear-cut thresholds to translate experimental readouts into binary test results ignores the impact of technical and biological variability on the precision of testing methods. Ignoring the precision of testing methods may lead to over- or underestimation errors regarding a testing method's predictive accuracy. Assessing the hazardous properties of substances for regulatory purposes requires, therefore, accurate as well as precise test results. This is an observation relevant for testing methods using *in vivo*, *in vitro* or *in silico* approaches. According to the findings in Chapter 4, the documentation of test results should be expanded and should include, besides positive (i.e. indicating an adverse effect) and negative (i.e. indicating no adverse effect) results, also information for which substances the test delivered inconclusive (i.e. borderline) results. If a substance is classified as borderline, further testing might become necessary in order to provide sufficient information for safety assessment of substances in a regulatory context. Quantifying the borderline range for several non-animal testing methods, underlines the need to revise the current way of evaluating and documenting testing methods' predictive accuracy. This is particularly relevant in a regulatory context, for example within the validation process of non-animal testing methods (Sauer et al., 2016).

Acknowledging that intra-assay variability of testing methods can affect their precision and their predictive accuracy stimulates reflection about the regulatory validation process and the criteria for evaluating non-animal testing methods. Specifically, our findings point to the need to consider quantitative estimates of a testing method's precision as an important piece of information in the evaluation of a method's predictive performance. Besides technical and biological variability it is important to consider factors affecting the intra- and inter-laboratory reproducibility of tests for example the absence of robust reference data for comparisons between test results (Hothorn, 2002; Hothorn, 2003). Our suggestion to describe predictive accuracy metrics using ranges, indicating the expected values of accuracy metrics,

rather than with point estimates, contributes to the ongoing discussion on how to assess this uncertainty, and the implications for evaluating the relative performance of animal tests in comparison to non-animal methods. Comparing information about testing methods' predictive accuracy across different testing methods should be based on a transparent documentation of this uncertainty, e.g. within test guidelines and validation reports.

6.3 Limitations of the thesis

In the following paragraphs we reflect on limitations with regard to the scope of this thesis, the assumptions underlying to the methodologies used, and data availability.

6.3.1 Scope

This thesis focused on a specific toxicological endpoint, i.e. skin sensitisation. Though being a highly relevant endpoint for safety assessments of chemicals under different regulatory frameworks, and although the findings about the features of optimising sequential testing strategies are generic, the applicability of the Bayesian VOI approach should be explored for different toxicological endpoints. In particular, the expected welfare gains and losses from releasing or banning substances used as cosmetic ingredients are highly case specific. Moreover, while for human health endpoints such as skin sensitisation, but also carcinogenicity, estimates of expected health damage costs can be retrieved from the scientific literature, this is more difficult for endpoints such as liver toxicity. Estimates of the (monetary) damage costs for environmental endpoints (e.g. aquatic toxicity) usually do not exist. Here, revealed or stated preference methods could be used for generating data on environmental externalities caused by a release of hazardous substances. Furthermore, the presented approach refers to individual substances only and did not account for additive or synergistic effects of chemicals which are part of mixtures. Though being beyond the scope of this thesis, developing optimised testing strategies for mixtures is an interesting option for further research.

The choice of skin sensitisation as endpoint for assessing the impact of biological and technical variability of testing methods (Chapter 4) and the uncertainties underlying to their predictivity (Chapter 5) was for illustrative purposes. Assessing the impact of intra-test variability on the precision of testing methods is a complex field of research which also has implications with regard to the policy-driven process of test validation. Addressing in detail different sources of uncertainties underlying to the precision and predictivity of testing methods goes beyond the scope of this thesis. Nonetheless, our findings underline that accuracy metrics should rather be documented as ranges rather than as point estimates. Using

test accuracy metrics as point estimates is a limitation which also applies to the scope of the decision-theoretic VOI analysis presented in Chapter 3. The VOI model could, therefore, be further expanded by integrating the uncertainty of test information due to the limited precision of testing methods into the Bayesian VOI model.

6.3.2 Assumptions

The quantitative methods presented in this thesis were based on a number of assumptions. For instance, to explore the applicability of the Bayesian VOI model presented in Chapter 3, Kathon CG was used as a proof-of-concept case. To quantify the expected payoffs from marketing Kathon CG, we made simplifying assumptions about the substance's marketing volume, its market price, and the slope of the supply and demand curves of this chemical. Clearly, modifying these assumptions impacts the expected value of information of individual non-animal testing methods and that of combinations of these methods into sequential or battery strategies. Although the numerical values of the analysis have to be treated with care, the VOI model can be straightforwardly used to investigate the impact of modified parameter values if better information, e.g. about the marketing volume of a substance, becomes available.

In addition, the Bayesian VOI model assumes risk neutral decision-makers. As a consequence, marginal health damage costs are considered to be constant. Given the increasing trend of skin allergies worldwide, particularly in children (Jackson et al., 2013), this assumption may not be valid and may not appropriately describe risk preferences of decision-makers in regulatory agencies. However, little information about decision-makers' risk preferences and risk perceptions regarding the use of specific groups of chemicals has become available so far. Further, the Bayesian VOI model was applied to a set of non-animal testing methods which are assumed to capture all key events in the skin sensitisation AOP. This implies that different key events have equal relevance and that information on different key events is of same importance. Relaxing this assumption could change the expected value of test information of individual testing methods, and could change the optimal order of testing methods in a sequence. In addition, for several toxicological endpoints (e.g. liver toxicity, see Landesmann et al., (2013); Vinken, (2013); Perkins et al., (2015)) existing knowledge of the events in the AOP is still rudimentary.

Furthermore, analysing the precision and the uncertainty of accuracy metrics of non-animal testing methods addresses the informative value of test outcomes. The analysis of the precision and implications of testing methods' precision constraints on the assessment of the predictive accuracy of non-animal testing methods was performed under the assumption that

the classification threshold is exogenously given. The approach for quantifying the borderline range (Chapter 4) around testing methods' classification threshold could, therefore, be generalised by accounting for varying thresholds. This might impact the borderline range, and consequently its impact on assessing the predictivity of a testing method. Further, our analysis on the impact of precision, sample size and sample composition on testing methods' accuracy assumes that the experimental samples are a representative selection of the entire set of chemicals. This assumption, which was underlying to the bootstrap analysis used, is difficult to verify. As a consequence, the outcomes of our analysis may not adequately cover the entire spectrum of substances.

6.3.3 Data availability

To quantify expected health damage costs from Kathon CG we used mean values for direct health treatment costs based on empirical estimates published in recent studies for different populations groups and in different European countries. Due to lacking data we did not account for potential health damage costs from occupational exposure or indirect costs such as impacts on the social life patients may suffer due to skin allergies (Both et al., 2007). Also, cost components such as a firm's loss of reputation due to erroneously marketing a skin sensitiser could not be included due to lacking information, e.g. about liability fees. For quantifying the number of people suffering from skin sensitisation in Europe, we used mean-estimates of skin sensitisation prevalence. While such estimates could be retrieved for Kathon CG, they are not available for many other cosmetic ingredients, and for substances relevant for endpoints other than skin sensitisation. Applying the Bayesian VOI model to any of these substances requires operating with mean prevalence estimates. This will impact the quantification of health damage costs and might affect the "true" optimal order of tests in a sequence. Further, key events of the AOP are considered to be equally relevant for causing an adverse outcome. Therefore, the non-animal testing methods(i.e. the DPRA, LuSens, the h-CLAT, and the OECD toolbox), which address different key events in the skin sensitisation AOP are treated with equal weights with respect to their relevance on addressing the skin sensitising properties of a substance. Finally, due to the lack of non-animal testing methods addressing the fourth key event in the skin sensitisation AOP, i.e. T-cell proliferation, the Bayesian VOI model as well as the analysis on the precision and accuracy of non-animal testing methods focused only on the first three key events in the AOP.

6.4 Conclusions

The main conclusions from this thesis can be summarised as follows:

1. The optimisation of testing strategies depends on the interplay of different parameters. These are a decision-maker's prior beliefs about the hazardous properties of a substance, non-animal testing method's predictive accuracy, social gains and losses corresponding to the set of decisions about the use of a substance, and testing costs.
2. The expected value of test information revealed from "gold standard" animal tests for assessing skin sensitisation is lower than that of individual non-animal methods. This result holds for the entire range of a decision maker's prior beliefs. When individual non-animal testing methods are combined into sequential or battery combinations, the expected value of test information increases.
3. Sequential or battery combinations of non-animal testing methods have higher expected value than standalone methods because a relative higher shift of posterior beliefs is observed when using testing methods in strategies. For strategies assessing skin sensitisation the increase of information gains exceeded the increase of testing costs.
4. The order of key events in the skin sensitisation AOP is neither a necessary nor a sufficient condition for optimising sequential testing. Furthermore, for a sequential testing strategy to be efficient not all key events in the AOP need necessarily to be covered.
5. Integrating information from non-animal testing methods into sequential strategies decreases the likelihood to misclassify substances due to the impact of biological and technical variability of individual testing methods.
6. Using clear-cut classification thresholds for transforming experimental readouts from testing methods into binary "hazardous/non-hazardous" information may lead to ambiguous classifications of substances. Quantifying the area where such ambiguous results are likely to occur, thus the borderline range can be used to document a testing method's limited precision.
7. The borderline range around the classification threshold of a testing method can impact a testing method's predictive accuracy. The size and direction of this impact depends on how many of substances of the experimental samples yield test results within or outside the borderline range.
8. Assessing the predictive accuracy of testing methods should be complemented by information about the borderline range. For a coherent evaluation of testing methods'

accuracy, the uncertainty of accuracy metrics caused by the limited precision of a testing method, captured in the borderline range should be assessed.

6.5 Suggestions for further research

In this section we provide suggestions for further research reflecting on the scope, assumptions and data availability of this thesis:

Reflecting on the scope, further research should focus on:

1. The optimisation of non-animal testing strategies for those substances which are of a high expected social relevance. For these substances the value of attaining information from testing will be highest.
2. The influence of the limited precision of testing methods on the optimisation of non-animal testing strategies. Specifically, one would expect that over- and under-estimation errors of accuracy metrics can change the optimal sequence of testing methods for given prior beliefs about the true hazardous properties of substances.

Reflecting on the assumptions, further research should focus on:

3. The impact of different risk preferences of stakeholders on the optimisation of testing strategies. We suggest conducting empirical research using, for example, choice experiments, to understand the behavioural drivers of different stakeholders (risk assessors in industry, regulatory agencies and consumers) for controlling the risks of chemicals.
4. Expanding the Bayesian VOI model in order to consider information about the AOP and the relevance of each key event regarding the formation of an adverse effect. Furthermore, uncertainties about key events in an AOP need to be included.
5. Further assessment of the prediction models of testing methods using the Receiver Operating Characteristic (ROC) curve analysis with varying classification thresholds while accounting for the impact of biological and technical variability of the precision of testing methods thus considering the borderline range.

Reflecting on data used, further research should focus on:

6. Compiling a dataset of different types of costs related to testing, e.g. health damage costs for consumers or loss of reputation for producers. This research could also expand on occupational exposure to substance in order to capture direct and indirect health damages costs occurring at workplaces.

Appendix A

Experimental samples of substances used in the borderline range calculation

Where substance names could not be published for reasons of data confidentiality we numbered them consecutively.

Table A1: Substances in the sample for calculating the BR of the DPRA prediction model

| Chemical name | No. of runs | Test substance concentrations considered [mM] | Study year |
|--|-------------|---|------------|
| Ethylene glycole dimethacrylate* (positive control) | 211 | 0.5, 1, 5, 10, 50, 100 | 2014 |
| Substance 1 | 24 | 1, 10, 100 | 2015 |
| Substance 2 | 6 | 100 | 2015 |
| Substance 3 | 3 | 100 | 2015 |
| Substance 4 | 12 | 1, 5, 10, 100 | 2014 |
| Substance 5 | 9 | 1, 5, 10 | 2014 |
| Substance 6 | 12 | 1, 5, 10, 100 | 2014 |
| Substance 7 | 12 | 1, 5, 10, 100 | 2014 |
| Substance 8 | 3 | 100 | 2014 |
| Substance 9 | 6 | 1, 10, 100 | 2015 |
| Substance 10 | 6 | 1, 10, 100 | 2015 |
| Substance 11 | 6 | 1, 10, 100 | 2015 |
| Substance 12 | 4 | 1, 10 | 2015 |
| Substance 13 | 6 | 1, 10, 100 | 2015 |
| Substance 14 | 6 | 1, 10, 100 | 2015 |
| Substance 15 | 4 | 1, 10 | 2015 |
| Substance 16 | 4 | 1, 10 | 2015 |
| Substance 17 | 6 | 1, 10, 100 | 2015 |
| Substance 18 | 6 | 1, 10, 100 | 2015 |
| Substance 19 | 4 | 1, 10 | 2014 |
| Substance 20 | 6 | 1, 10, 100 | 2014 |
| Substance 21 | 6 | 1, 10, 100 | 2014 |
| Substance 22 | 6 | 1, 10, 100 | 2014 |
| Substance 23 | 6 | 1, 10, 100 | 2014 |
| Substance 24 | 6 | 1, 10, 100 | 2014 |
| Substance 25 | 6 | 1, 10, 100 | 2014 |
| Substance 26 | 6 | 1, 10, 100 | 2014 |
| Substance 27 | 6 | 1, 10, 100 | 2014 |
| Substance 28 | 6 | 1, 10, 100 | 2014 |
| Substance 29 | 6 | 1, 10, 100 | 2014 |
| Substance 30 | 3 | 100 | 2014 |
| Substance 31 | 3 | 100 | 2014 |
| Substance 32 | 3 | 100 | 2014 |
| Substance 33 | 3 | 100 | 2014 |
| Substance 34 | 3 | 100 | 2013 |
| Substance 35 | 3 | 3.76% | 2013 |
| Substance 36 | 3 | 3.76% | 2013 |
| Substance 37 | 3 | 100 | 2013 |
| Substance 38 | 3 | 100 | 2013 |
| Substance 39 | 3 | 100 | 2013 |
| Substance 40 | 3 | 100 | 2012 |
| Substance 41 | 3 | 100 | 2013 |

Table A2: Substances in the sample for calculating the BR of the LuSens prediction model

| Chemical name | No. of runs | Test substance range of concentrations considered [µg/mL] | Study Year |
|--------------------------------|-------------|---|------------|
| Lactic Acid (Negative control) | 395 | 450 | 2013-2015 |
| Substance 1 | 36 | 1.4 - 5.1 | 2013-2015 |
| Substance 2 | 53 | 0.15 - 0.54 | 2013-2015 |
| Substance 3 | 48 | 3.9 - 14.0 | 2013-2015 |
| Substance 4 | 144 | 0.06 - 0.92 | 2013-2015 |
| Substance 5 | 174 | 0.65 - 16.8 | 2013-2015 |
| Substance 6 | 96 | 77.0 - 479.0 | 2013-2015 |
| Substance 7 | 96 | 0.03 - 0.53 | 2013-2015 |
| Substance 8 | 96 | 0.31 - 4.75 | 2013-2015 |
| Substance 9 | 48 | 4.5 - 16.2 | 2013-2015 |
| Substance 10 | 48 | 4.7 - 42.9 | 2013-2015 |
| Substance 11 | 120 | 4.0 - 176.0 | 2013-2015 |
| Substance 12 | 48 | 6.0 - 46.0 | 2013-2015 |
| Substance 13 | 72 | 3.9 - 20.2 | 2013-2015 |
| Substance 14 | 48 | 4.6 - 16.5 | 2013-2015 |
| Substance 15 | 48 | 0.44 - 1.57 | 2013-2015 |
| Substance 16 | 48 | 1283.0 - 7942.0 | 2013-2015 |
| Substance 17 | 48 | 636.0 - 2278.0 | 2013-2015 |
| Substance 18 | 48 | 14.0 - 85.0 | 2013-2015 |
| Substance 19 | 48 | 226.0 - 2012.0 | 2013-2015 |
| Substance 20 | 36 | 0.59 - 2.02 | 2013-2015 |
| Substance 21 | 72 | 0.91 - 3.27 | 2013-2015 |
| Substance 22 | 96 | 6.9 - 106.0 | 2013-2015 |
| Substance 23 | 96 | 1.8 - 28.0 | 2013-2015 |
| Substance 24 | 72 | 542.0 - 1941.0 | 2013-2015 |
| Substance 25 | 72 | 450.0 - 2000.0 | 2013-2015 |

Table A3: Substances in the sample for calculating the BR of the h-CLAT prediction model

| Chemical name | No. of runs | Test substance concentrations considered [µg/mL] | Study Year |
|--------------------------------|-------------|--|------------|
| Lactic Acid (Negative control) | 53 | 1000 | 2013-2015 |
| DNCB (Positive control) | 53 | 4.0 | 2013-2015 |
| Substance 1 | 32 | 7396.0 - 2064.0 | 2013-2015 |
| Substance 2 | 30 | 5655.0 - 29176.0 | 2013-2015 |
| Substance 3 | 40 | 12.3 - 64.0 | 2013-2015 |
| Substance 4 | 32 | 69.0 - 510.0 | 2013-2015 |
| Substance 5 | 48 | 14.2 - 73.4 | 2013-2015 |
| Substance 6 | 48 | 2.2 - 8.0 | 2013-2015 |
| Substance 7 | 32 | 29.0 - 105.0 | 2013-2015 |
| Substance 8 | 48 | 1589.0 - 5695.0 | 2013-2015 |
| Substance 9 | 48 | 13.0 - 48.0 | 2013-2015 |
| Substance 10 | 32 | 115.0 - 710.0 | 2013-2015 |
| Substance 11 | 32 | 14.0 - 51.0 | 2013-2015 |

Table A4: Substances in the sample for calculating the BR of the LLNA prediction model

| Chemical name | CAS no | No. of runs | Test substance concentrations considered [$\mu\text{g/mL}$] | Study Year |
|------------------------------------|---------------------------|-------------|---|------------|
| DL-Lactic acid | 50-21-5 | 20 | 5, 10, 25 | 2013-2015 |
| Salicylic acid | 69-72-7 | 20 | 5, 10, 25 | 2013-2015 |
| Chlorobenzene | 108-90-7 | 20 | 25, 50, 100 | 2013-2015 |
| Methyl methacrylate | 80-62-6 | 20 | 25, 50, 100 | 2013-2015 |
| 2-Mercaptobenzothiazole | 149-30-4 | 19 | 0.75, 2.0, 7.5 | 2013-2015 |
| Methyl salicylate | 119-36-8 | 20 | 10, 25, 50 | 2013-2015 |
| MCI / MI | 26172-55-4 & 2682-20-4 | 40 | 0.005, 0.05, 0.2, 0.1, 0.5, 1 | 2013-2015 |
| Sodium dodecyl sulfate | 151-21-3 | 20 | 1, 5, 10 | 2013-2015 |
| Imidazolidinyl urea | 39236-46-9 | 20 | 10, 25, 50 | 2013-2015 |
| Ethylenglycolmethacrylate (EGDMA) | 97-90-5 | 20 | 25, 50, 100 | 2013-2015 |
| Nickel(II) chloride | 7718-54-9 | 20 | 1, 2.5, 5 | 2013-2015 |
| Cinnamic alcohol | 104-54-1 | 20 | 10, 25, 50 | 2013-2015 |
| Isopropanol | 67-63-0 | 20 | 25, 50, 100 | 2013-2015 |
| Phenylbenzoate | 93-99-2 | 40 | 5, 10, 15, 25, 40 | 2013-2015 |
| Isoeugenol | 97-54-1 | 20 | 1, 5, 10 | 2013-2015 |
| Xylene | 1330-20-7 | 20 | 25, 50, 100 | 2013-2015 |
| Alpha-Hexylcinnamaldehyde | 101-86-0 | 20 | 5, 10, 25 | 2013-2015 |
| p-Phenylenediamine | 106-50-3 | 20 | 0.05, 0.1, 0.5 | 2013-2015 |
| Citral | 5392-40-5 | 20 | 5, 10, 25 | 2013-2015 |
| Cobalt(II) chloride | 7646-79-9 | 20 | 0.25, 0.5, 1 | 2013-2015 |
| Eugenol | 97-53-0 | 20 | 2.5, 10, 25 | 2013-2015 |
| 1-Chloro-2,4-dinitrobenzene (DNCB) | 97-00-7 | 20 | 0.025, 0.1, 0.25 | 2013-2015 |

Appendix B

Experimental samples used for identifying borderline substances and calculate the predictive accuracy of non-animal testing methods (with including and excluding the borderline substances) when compared to the animal test LLNA and to the human data as reference

Table B1: Experimental sample tested with the DPRA (borderline substances in bold)

| Chemical name | CAS no | Sensitisation potential1 in mice or humans (by conventional approach, assessed without <i>BR</i>) | |
|---|------------|--|--------------|
| | | LLNA | Human |
| Salicylic acid | 69-72-7 | N | N |
| Geraniol | 106-24-1 | P | - |
| Benzyl alcohol | 100-51-6 | P | P |
| Tween 80 | 9005-65-6 | N | P |
| 3-Dimethylamino propylamine | 109-55-7 | N | N |
| cis-6-Nonenal | 2277-19-2 | P | P |
| Ethyl vanillin | 121-32-4 | P | - |
| Undecylenic acid | 112-38-9 | N | - |
| 2-Methoxy-4-Methylphenol | 93-51-6 | P | P |
| Ethyl benzoylacetate | 94-02-0 | P | - |
| Dihydroeugenol (2-Methoxy-4-Propyl-phenol) | 2785-87-7 | N | - |
| α -Hexyl cinnamic aldehyde | 101-86-0 | P | - |
| N,N-Diethyl-m-toluanimide | 134-62-3 | N | - |
| Penicillin G | 61-33-6 | P | P |
| d,l-Citronellol | 106-22-9 | P | N |
| Pentachlorophenol | 87-86-5 | P | P |
| p-tert-Butyl-alpha-ethyl hydrocinnamal (Lilial) | 80-54-6 | P | P |
| 1-Bromobutane | 109-65-9 | N | - |
| Fumaric acid | 110-17-8 | N | N |
| Glucose | 50-99-7 | N | N |
| Propyl paraben | 94-13-3 | TN | TN |
| 4-Methoxyacetophenone (Acetanisole) | 100-06-1 | TN | TN |
| 6-Methylcoumarin | 92-48-8 | TN | TN |
| Nonanoic acid | 112-05-0 | FN | TN |
| Isopropanol | 67-63-0 | FN | TN |
| Methyl salicylate | 119-36-8 | TN | TN |
| Dibutyl phthalate | 84-74-2 | TN | inconclusive |
| Pyridine | 110-86-1 | TN | - |
| dl- α -Tocopherol | 10191-41-0 | TN | - |
| Clotrimazole | 23593-75-1 | FN | - |
| Methyl pyruvate | 600-22-6 | FN | - |
| 1-Butanol | 71-36-3 | TN | TN |
| Xylene | 1330-20-7 | FN | TN |
| Diethyl phthalate | 84-66-2 | TN | TN |
| Vinylidene dichloride | 75-35-4 | TN | - |
| Oxalic acid anhydrous | 144-62-7 | FN | - |
| Octanoic acid (Caprylic acid) | 124-07-2 | TN | TN |
| Coumarin | 91-64-5 | TN | FN |
| Dimethyl formamide | 68-12-2 | TN | - |
| Glycerol | 56-81-5 | TN | TN |
| 2,2,6,6-Tetramethyl-3,5-heptanedione | 1118-71-4 | FN | - |
| N,N-Dibutylaniline | 613-29-6 | FN | FN |
| Resorcinol | 108-46-3 | FN | - |
| Chlorobenzene | 108-90-7 | TN | - |
| Propylene glycol (1,2-Propanediol) | 57-55-6 | TN | TN |
| 4-Chloroaniline | 106-47-8 | FN | - |
| 7,12-Dimethylbenz[α]anthracene | 57-97-6 | FN | - |

| | | | |
|---|-------------|----|--------------|
| Aniline | 62-53-3 | FN | FN |
| Saccharin | 81-07-2 | TN | TN |
| Hexadecyltrimethylammonium bromide (Cetrimide) | 57-09-0 | TN | - |
| n-Hexane | 110-54-3 | TN | TN |
| Benzalkonium chloride | 8001-54-5 | TN | TN |
| Lactic acid | 50-21-5 | TN | TN |
| Octanenitrile | 124-12-9 | TN | - |
| Undec-10-enal | 112-45-8 | FN | - |
| Benzyl benzoate | 120-51-4 | FN | TN |
| Methyl 4-hydroxybenzoate (Methylparaben) | 99-76-3 | TN | - |
| Butylbenzylphthalate | 85-68-7 | TN | - |
| 4-Hydroxybenzoic acid | 99-96-7 | TN | TN |
| Sulfanilamide | 63-74-1 | TN | TN |
| Cocamidopropyl betaine | 61789-40-0 | TN | - |
| Benzene,1-methoxy-4-methyl-2-nitro (4-Methyl-2-nitroanisole) | 119-10-8 | TN | - |
| Squaric acid diethyl ester | 5231-87-8 | FN | - |
| Clofibrate (Ethyl (2-(4-chlorophenoxy)-2-methylpropanoate) | 637-07-0 | TN | - |
| α -Amyl cinnamic aldehyde | 122-40-7 | FN | Inconclusive |
| Streptomycin sulfate | 3810-74-0 | TN | FN |
| α -iso-Methylionone | 127-51-5 | FN | TN |
| Carbonic acid, dioctyl ester | 1680-31-5 | TN | - |
| Hexyl salicylate | 6259-76-3 | FN | TN |
| Benzyl cinnamate | 103-41-3 | FN | - |
| Benzyl salicylate | 118-58-1 | FN | TN |
| Sulfanilic acid | 121-57-3 | TN | - |
| Isopropyl myristate a | 110-27-0 | FN | TN |
| p-Aminobenzoic acid | 150-13-0 | TN | TN |
| Tartaric acid | 87-69-4 | TN | TN |
| Zinc sulfate | 7733-02-0 | TN | - |
| Dioctyl ether | 629-82-3 | TN | - |
| 2,2-Azobis phenol | 2050-14-8 | FN | - |
| Benzaldehyde | 100-52-7 | TN | FN |
| Farnesol | 4602-84-0 | FN | FN |
| 3-Aminophenol | 591-27-5 | FN | - |
| (+/-) Linalool | 78-70-6 | FN | TN |
| Diethylenetriamine | 111-40-0 | FN | FN |
| Octanoic acid, 4-methyl-2-pentylbutyl ester | 868839-23-0 | TN | - |
| R(+)-Limonene | 5989-27-5 | TP | FP |
| Ethylenediamine free base | 107-15-3 | TP | TP |
| Vanillin | 121-33-5 | FP | FP |
| Cyclamen aldehyde | 103-95-7 | TP | - |
| Tropolone | 533-75-5 | TP | - |
| Cinnamyl Alcohol | 104-54-1 | TP | TP |
| R-Carvone | 6485-40-1 | TP | TP |
| Benzocaine | 94-09-7 | FP | TP |
| 3-Phenoxypropiononitrile | 3055-86-5 | FP | - |
| 2-Acetyl-cyclohexanone | 874-23-7 | FP | - |
| Diethyl sulfate | 64-67-5 | TP | - |
| 2-Phenylpropionaldehyde | 93-53-8 | TP | TP |
| 5-Methyl-2,3-hexanedione | 13706-86-0 | TP | TP |
| 1-Iodoheptane | 638-45-9 | FP | - |
| 2,2-Bis-[4-(2-hydroxy-3-methacryloxypropoxy)phenyl]-propane (Bis-GMA) | 1565-94-2 | TP | - |
| Farnesal | 502-67-0 | TP | - |
| α -Methyl-trans-Cinnamaldehyde | 101-39-3 | TP | - |
| 3,4-Dihydrocoumarin | 119-84-6 | TP | TP |
| Eugenol | 97-53-0 | TP | TP |
| Lyril / 3 and 4-(4-Hydroxy-4-methylpentyl)-3-cyclohexene-1-carboxaldehyde | 31906-04-4 | TP | TP |
| Nickel chloride | 7718-54-9 | FP | TP |
| Bisphenol A-diglycidyl ether | 1675-54-3 | TP | TP |
| 1,2,4-Benzenetricarboxylic anhydride (Trimellitic anhydride) | 552-30-7 | TP | - |
| 1-(p-Methoxyphenyl)-1-penten-3-one | 104-27-8 | TP | - |
| 3-Propylenediphenylphthalide | 17369-59-4 | TP | TP |
| Perillaldehyde | 2111-75-3 | TP | TP |

Appendix

| | | | |
|---|------------------------|----|----|
| Tetrachloro-salicylanilide | 1154-59-2 | TP | TP |
| 2-Fluoro-5-nitroaniline | 369-36-8 | FP | - |
| Phthalic anhydride | 85-44-9 | TP | FP |
| 1,2-cyclohexane dicarboxylic anhydride | 85-42-7 | TP | - |
| Squaric acid | 2892-51-5 | TP | TP |
| Formaldehyde | 50-00-0 | TP | TP |
| 2-Hydroxypropyl methacrylate | 923-26-2 | FP | - |
| 1-Phenyl-1,2-propanedione | 579-07-7 | TP | - |
| Cobalt chloride | 7646-79-9 | TP | TP |
| Methylmethacrylate | 80-62-6 | TP | TP |
| Phenyl benzoate | 93-99-2 | TP | TP |
| 3-Chloro-4-Methoxybenzaldehyde (3-Chloro-p-anisaldehyde) | 4903-09-7 | FP | - |
| Butyl glycidyl ether | 2426-08-6 | TP | TP |
| Imidazolidinyl urea | 39236-46-9 | TP | TP |
| 1-Naphthol | 90-15-3 | TP | - |
| Ethanol-2-butoxy acetate | 112-07-2 | FP | - |
| 1-Bromohexane | 111-25-1 | TP | - |
| Phenylacetaldehyde | 122-78-1 | TP | TP |
| Benzoic acid | 65-85-0 | FP | - |
| 1-Iodohexadecane | 544-77-4 | TP | - |
| Citral | 5392-40-5 | TP | TP |
| Bandrowski's Base (N,N-bis(4-aminophenyl)-2,5-diamino-1,4-quinone-diimine) | 20048-27-5 | TP | - |
| 1,1,3-Trimethyl-2-Formylcyclohexa-2,4-diene (Safranal) | 116-26-7 | TP | TP |
| 4-Vinyl pyridine | 100-43-6 | TP | - |
| Benzylidene acetone (4-Phenyl-3-buten-2-one) | 122-57-6 | TP | TP |
| 2-Nitro-1,4-phenylenediamine | 5307-14-2 | TP | TP |
| 2,5-Diaminotoluene sulfate (PTD) | 615-50-9 | TP | TP |
| Hydroxycitronellal | 107-75-5 | TP | TP |
| MCI/MI | 26172-55-4 & 2682-20-4 | TP | TP |
| Sodium lauryl sulfate / sodium dodecyl sulfate (SDS) | 151-21-3 | TP | FP |
| Methyl-2-octynoate / Methyl heptene carbonate | 111-12-6 | TP | TP |
| 2-Methyl-2H-Isothiazol-3-one (MI) | 2682-20-4 | TP | TP |
| 4-Allylanisole | 140-67-0 | TP | - |
| Diphenylcyclopropanone | 886-38-4 | TP | TP |
| Lauryl gallate | 1166-52-5 | TP | TP |
| Iodopropynyl butylcarbamate | 55406-53-6 | TP | TP |
| Furil | 492-94-4 | FP | - |
| 2-Methylundecanal | 110-41-8 | TP | - |
| N,N-dimethyl-4-nitrosoaniline | 138-89-6 | TP | - |
| 2-Propylheptyl acrylate | 149021-58-9 | TP | - |
| trans-2-Hexenal | 6728-26-3 | TP | TP |
| 5-Amino-2-methylphenol | 2835-95-2 | TP | - |
| Chlorothalonil | 1897-45-6 | TP | - |
| 2-Mercaptobenzothiazole | 149-30-4 | TP | TP |
| Methyl 2-nonynoate | 111-80-8 | TP | TP |
| Methyl methanesulphonate | 66-27-3 | TP | - |
| 4-(N-Ethyl-N-2-methan-sulphonamido-ethyl)-2-methyl-1,4-phenylenediamine (CD3) | 25646-71-3 | TP | - |
| 1,2-Dibromo-2,4-dicyanobutane (MDGN, Methylidibromo glutaronitrile) | 35691-65-7 | TP | TP |
| Trans-2-Decenal | 3913-71-1 | TP | - |
| Tetramethylthiuram disulfide | 137-26-8 | TP | TP |
| 1,2-Benzisothiazolin-3-One (Proxel active) | 2634-33-5 | TP | TP |
| Propanoic acid, 3-Bromo-Mmethyl ester (Methyl-3-bromopropionate) | 3395-91-3 | FP | - |
| 4-Carboxyphenylacetate | 2345-34-8 | TP | - |
| Cinnamic aldehyde | 104-55-2 | TP | TP |
| 2-Aminophenol | 95-55-6 | TP | TP |
| Diethyl acetaldehyde | 97-96-1 | TP | - |
| Glutaraldehyde | 111-30-8 | TP | TP |
| Abietic acid | 514-10-3 | TP | TP |
| 4-Ethoxymethylene-2-phenyl-2-oxazolin-5-one (Oxazolone) | 15646-46-5 | TP | TP |
| 4-Amino-m-cresol | 2835-99-6 | TP | - |

| | | | |
|--|------------|----|----|
| Isoeugenol | 97-54-1 | TP | TP |
| 2-Ethylhexyl acrylate | 103-11-7 | TP | - |
| 2,4-Heptadienal | 5910-85-0 | TP | - |
| 2,4-Dinitrobenzenesulfonic acid, sodium salt | 885-62-1 | TP | - |
| Benzyl bromide | 100-39-0 | TP | - |
| 2,4,6-Trinitrobenzenesulfonic acid | 2508-19-2 | TP | - |
| Propyl gallate | 121-79-9 | TP | TP |
| 4-Nitrobenzyl bromide | 100-11-8 | TP | - |
| Glyoxal | 107-22-2 | TP | TP |
| Ethylene glycol dimethacrylate (EGDMA) | 97-90-5 | TP | TP |
| 2,3-Butanedione | 431-03-8 | TP | - |
| Isophorone diisocyanate | 4098-71-9 | TP | - |
| 5-Chloro-2-methyl-4-isothiazolin-3-one (MCI) | 26172-55-4 | TP | - |
| 1,6-hexamethylene diisocyanate | 822-06-0 | TP | - |
| Hydroquinone | 123-31-9 | TP | TP |
| Maleic anhydride | 108-31-6 | TP | - |
| 1,4-Phenylenediamine | 106-50-3 | TP | TP |
| 4-(Methylamino) Phenol sulfate (Metol) | 55-55-0 | TP | TP |
| 1-Chloro-2,4-Dinitrobenzene (Dinitrochlorobenzene, DNCB) | 97-00-7 | TP | TP |
| Fluorescein-5-isothiocyanate | 3326-32-7 | TP | - |
| 3-Methylcatechol | 488-17-5 | TP | - |
| Diethyl maleate | 141-05-9 | TP | TP |
| Benzoyl peroxide | 94-36-0 | TP | TP |
| 2-Hydroxyethyl acrylate | 818-61-1 | TP | TP |
| Ethyl acrylate | 140-88-5 | TP | TP |
| Methyl acrylate | 96-33-3 | TP | - |
| Butyl acrylate | 141-32-2 | TP | - |
| p-Benzoquinone | 106-51-4 | TP | TP |
| Tosylchloramide sodium (Chloramine T) | 127-65-1 | TP | - |

Table B2: Experimental sample tested with LuSens(borderline substances in bold)

| Chemical name | CAS no | Sensitisation potential1 in mice or humans (by conventional approach, assessed without BR) | |
|--|-------------------|--|----------|
| | | LLNA | Human |
| 1-Butanol | 71-36-3 | N | N |
| Benzoyl peroxide | 94-36-0 | P | P |
| 4-Allylanisole | 140-67-0 | P | - |
| Methyldibromo glutaronitrile (MDGN) | 35691-65-7 | P | P |
| Phthalic anhydride | 85-44-9 | FN | TN |
| Resorcinol | 108-46-3 | FN | FN |
| Sodium lauryl sulfate / sodium dodecyl sulfate (SDS) | 151-21-3 | FN | TN |
| Nickel chloride | 7718-54-9 | TN | FN |
| Salicylic acid | 69-72-7 | TN | TN |
| Farnesal | 502-67-0 | FN | - |
| Propyl gallate | 121-79-9 | FN | FN |
| Hexadecyltrimethylammonium bromide (Cetrimide) | 57-09-0 | TN | TN |
| Lactic acid | 50-21-5 | TN | TN |
| Aniline | 62-53-3 | FN | FN |
| 4-Hydroxybenzoic acid | 99-96-7 | TN | TN |
| Glucose | 50-99-7 | TN | TN |
| Sulfanilamide | 63-74-1 | TN | TN |
| Penicillin G | 61-33-6 | FN | FN |
| p-Aminobenzoic acid | 150-13-0 | TN | TN |
| Ethylenediamine free base | 107-15-3 | FN | FN |
| Phenyl benzoate | 93-99-2 | FN | FN |
| Glycerol | 56-81-5 | TN | TN |
| Cocamidopropyl betaine | 61789-40-0 | TN | - |
| Propylene glycol (1,2-Propanediol) | 57-55-6 | TN | TN |
| n-Hexane | 110-54-3 | TN | TN |
| Isopropanol | 67-63-0 | TN | TN |
| Fumaric acid | 110-17-8 | TN | TN |
| Tartaric acid | 87-69-4 | TN | TN |
| Xylene | 1330-20-7 | FN | TN |

Appendix

| | | | |
|---|---------------------------|----|--------------|
| Pyridine | 110-86-1 | FN | TN |
| Vanillin | 121-33-5 | TN | TN |
| Octanoic acid, 4-methyl-2-pentylbutyl ester | 868839-23-0 | TN | - |
| Benzyl alcohol | 100-51-6 | FP | TP |
| Dioctyl ether | 629-82-3 | FP | - |
| Hydroxycitronellal | 107-75-5 | TP | TP |
| Methyl salicylate | 119-36-8 | FP | FP |
| 1,6-hexamethylene diisocyanate | 822-06-0 | TP | - |
| p-Benzoquinone | 106-51-4 | TP | TP |
| Potassium dichromate | 7778-50-9 | TP | TP |
| 4-Nitrobenzyl bromide | 100-11-8 | TP | - |
| α-Hexyl cinnamic aldehyde | 101-86-0 | TP | inconclusive |
| 1-Chloro-2,4-dinitrobenzene (Dinitrochlorobenzene, DNCB) | 97-00-7 | TP | TP |
| Diethyl phthalate | 84-66-2 | TN | TN |
| 2-Ethylhexyl acrylate | 103-11-7 | TP | - |
| 2-Phenylpropionaldehyde | 93-53-8 | TP | TP |
| 6-Methylcoumarin | 92-48-8 | FP | inconclusive |
| Tween 80 | 9005-65-6 | FP | FP |
| Propyl paraben (propyl-4-hydroxybenzoate) | 94-13-3 | FP | FP |
| Formaldehyde | 50-00-0 | TP | TP |
| Isophorone diisocyanate | 4098-71-9 | TP | - |
| 2-Propylheptyl acrylate | 149021-58-9 | TP | - |
| Glyoxal | 107-22-2 | TP | TP |
| Ethyl acrylate | 140-88-5 | TP | TP |
| Imidazolidinyl urea | 39236-46-9 | TP | TP |
| Butyl glycidyl ether | 2426-08-6 | TP | TP |
| Tetramethylthiuram disulfide | 137-26-8 | TP | TP |
| Eugenol | 97-53-0 | TP | TP |
| 2,4,6-Trinitrobenzenesulfonic acid | 2508-19-2 | TP | - |
| Glutaraldehyde | 111-30-8 | TP | TP |
| Methyl 4-hydroxybenzoate (Methylparaben) | 99-76-3 | FP | - |
| MCI/MI | 26172-55-4 & 2682-20-4 | TP | TP |
| Cinnamyl Alcohol | 104-54-1 | TP | TP |
| Methylmethacrylate | 80-62-6 | TP | TP |
| Cobalt chloride | 7646-79-9 | TP | TP |
| 4-Ethoxymethylene-2-phenyl-2-oxazolin-5-one (Oxazolone) | 15646-46-5 | TP | TP |
| 4-(Methylamino)phenol sulfate (Metol) | 55-55-0 | TP | TP |
| Undecylenic acid | 112-38-9 | TP | TP |
| 2,3-Butanedione | 431-03-8 | TP | - |
| 4-Methoxyacetophenone (Acetanisole) | 100-06-1 | FP | FP |
| Butyl acrylate | 141-32-2 | TP | - |
| 1,4-Phenylenediamine | 106-50-3 | TP | TP |
| Methyl acrylate | 96-33-3 | TP | - |
| Diethyl maleate | 141-05-9 | TP | TP |
| Benzylidene acetone (4-Phenyl-3-buten-2-one) | 122-57-6 | TP | TP |
| Cinnamic aldehyde | 104-55-2 | TP | TP |
| 2-Mercaptobenzothiazole | 149-30-4 | TP | TP |
| Isoeugenol | 97-54-1 | TP | TP |
| Ethylene glycol dimethacrylate (EGDMA) | 97-90-5 | TP | TP |
| Citral | 5392-40-5 | TP | TP |

Table B3: Experimental sample tested with the h-CLAT(borderline substances in bold)

| Chemical name | CAS no | Sensitisation potential1 in mice or humans (by conventional approach, assessed without <i>BR</i>) | |
|--|------------------------|--|-------|
| | | LLNA | Human |
| 4-phenylenediamine | 106-50-3 | P | P |
| Phenyl benzoate | 93-99-2 | P | P |
| Ethylene diamine | 107-15-3 | P | P |
| Aniline | 62-53-3 | P | P |
| Farnesal | 502-67-0 | P | - |
| Methyldibromo Glutaronitrile (MDGN) | 35691-65-7 | P | P |
| p-Benzoquinone | 106-51-4 | P | P |
| Propyl gallate | 121-79-9 | P | P |
| MCI/MI | 26172-55-4 & 2682-20-4 | TP | TP |
| 1-chloro-2,4-dinitrobenzene | 97-00-7 | TP | - |
| Cobalt chloride | 7646-79-9 | TP | TP |
| Citral | 5392-40-5 | TP | TP |
| Cinnamic alcohol | 104-54-1 | TP | TP |
| Methylmethacrylate | 80-62-6 | TP | TP |
| Isopropanol | 67-63-0 | TN | TN |
| DL-lactic acid | 50-21-5 | TN | TN |
| Methyl salicylate | 119-36-8 | TN | TN |
| Sodium lauryl sulfate | 151-21-3 | FP | TN |
| Ethylene glycol dimethacrylate (EDGMA) | 97-90-5 | TP | TP |
| Xylene | 1330-20-7 | FN | TN |
| Sulfanilamide | 63-74-1 | TN | TN |
| 2,4,6-trinitrobenzenesulfonic acid | 2508-19-2 | FN | - |
| 2,3-butanedione | 431-03-8 | TP | - |
| 2-phenylpropionaldehyde | 93-53-8 | TP | TP |
| 4-allylanisole | 140-67-0 | TP | - |
| Benzylidene acetone | 122-57-6 | TP | TP |
| Diethyl maleate | 141-05-9 | TP | TP |
| Fumaric acid | 110-17-8 | TN | TN |
| Glucose | 50-99-7 | TN | TN |
| Hydroxycitronellal | 107-75-5 | TP | TP |
| p-aminobenzoic acid | 150-13-0 | TN | TN |
| Phthalic anhydride | 85-44-9 | FN | TP |
| Undecylenic acid | 112-38-9 | TP | TP |
| Vanillin | 121-33-5 | TN | TN |
| Propyl-4-hydroxybenzoate | 99-76-3 | FP | FP |
| Tartaric acid | 87-69-4 | TN | TN |
| n-hexane | 110-54-3 | TN | TN |
| Hexadecyltrimethylammonium bromid | 57-09-0 | TN | TN |
| Glycerol | 56-81-5 | TN | TN |
| Propylene glycol (1,2-Propanediol) | 57-55-6 | TN | TN |

Table B4: Experimental sample tested with the '2 out of 3'ITS approach (borderline substances in bold)

| Name | CAS no | Sensitisation potential ¹ in mice or humans (by conventional approach, assessed without <i>BR</i>) | |
|--|------------------------|--|----------|
| | | LLNA | Human |
| Phenyl benzoate | 93-99-2 | P | P |
| Ethylene diamine | 107-15-3 | P | P |
| Methyldibromo glutaronitrile (MDGN) | 35691-65-7 | P | P |
| Propyl gallate | 121-79-9 | P | P |
| Propylene glycol (1,2-Propanediol) | 57-55-6 | TN | TN |
| Tartaric acid | 87-69-4 | TN | TN |
| Glycerol | 56-81-5 | TN | TN |
| n-Hexane | 110-54-3 | TN | TN |
| Propyl paraben (Propyl-4-Hydroxybenzoate) | 99-76-3 | FP | FP |
| Sulfanilamide | 63-74-1 | TN | TN |
| Vanillin | 121-33-5 | TN | TN |
| Isopropanol | 67-63-0 | TN | TN |
| Lactic acid | 50-21-5 | TN | TN |
| Methyl salicylate | 119-36-8 | TN | TN |
| Fumaric acid | 110-17-8 | TN | TN |
| Glucose | 50-99-7 | TN | TN |
| p-Aminobenzoic acid | 150-13-0 | TN | TN |
| Hexadecyltrimethylammonium bromide (Cetrimide) | 57-09-0 | TN | TN |
| Xylene | 1330-20-7 | FN | TN |
| Methylmethacrylate | 80-62-6 | TP | TP |
| Aniline | 62-53-3 | FN | FN |
| Ethylene glycol dimethacrylate (EGDMA) | 97-90-5 | TP | TP |
| Undecylenic acid | 112-38-9 | TP | TP |
| Hydroxycitronellal | 107-75-5 | TP | TP |
| Cinnamyl Alcohol | 104-54-1 | TP | TP |
| 4-Allylanisole | 140-67-0 | TP | |
| Sodium lauryl sulfate / sodium dodecyl sulfate (SDS) | 151-21-3 | FN | TN |
| Farnesal | 502-67-0 | TP | |
| 2,3-Butanedione | 431-03-8 | TP | |
| Citral | 5392-40-5 | TP | TP |
| 2-Phenylpropionaldehyde | 93-53-8 | TP | TP |
| Benzylidene acetone (4-Phenyl-3-buten-2-one) | 122-57-6 | TP | TP |
| Diethyl maleate | 141-05-9 | TP | TP |
| Cobalt chloride | 7646-79-9 | TP | TP |
| 2,4,6-Trinitrobenzenesulfonic acid | 2508-19-2 | TP | |
| Phthalic anhydride | 85-44-9 | FN | TN |
| 1,4-Phenylenediamine | 106-50-3 | TP | TP |
| 1-Chloro-2,4-dinitrobenzene (Dinitrochlorobenzene, DNCB) | 97-00-7 | TP | TP |
| p-Benzoquinone | 106-51-4 | TP | TP |
| MCI/MI | 26172-55-4 & 2682-20-4 | TP | TP |

Table B5: Experimental sample tested with the LLNA (borderline substances in bold)

| Chemical name | CAS no | Sensitisation potential ¹ in mice or humans (by conventional approach, assessed without <i>BR</i>) | |
|------------------------------------|------------------------|---|-------|
| | | LLNA | Human |
| Salicylic acid | 69-72-7 | N | TN |
| Methyl methacrylate | 80-62-6 | P | P |
| Chlorobenzene | 108-90-7 | N | |
| Nickel chloride | 7718-54-9 | N | P |
| Phenyl benzoate | 93-99-2 | P | P |
| Methyl salicylate | 119-36-8 | N | TN |
| DL-Lactic acid | 50-21-5 | N | TN |
| 2-Mercaptobenzothiazole | 149-30-4 | P | TP |
| MCI / MI | 26172-55-4 & 2682-20-4 | P | TP |
| Sodium dodecyl sulfate | 151-21-3 | P | FN |
| Imidazolidinyl urea | 39236-46-9 | P | TP |
| Ethylenglycolmethacrylate (EGDMA) | 97-90-5 | P | TP |
| Cinnamic alcohol | 104-54-1 | P | TP |
| Isopropanol | 67-63-0 | N | TN |
| Isoeugenol | 97-54-1 | P | TP |
| Xylene | 1330-20-7 | P | FN |
| Alpha-Hexylcinnamaldehyde | 101-86-0 | P | TP |
| p-Phenylenediamine | 106-50-3 | P | - |
| Citral | 5392-40-5 | P | TP |
| Cobalt(II) chloride | 7646-79-9 | P | TP |
| Eugenol | 97-53-0 | P | TP |
| 1-Chloro-2,4-dinitrobenzene (DNCB) | 97-00-7 | P | TP |

Appendix C

Distributions of accuracy metrics for non-animal testing methods and the “2 out of 3” ITS - Using the LLNA as reference test
DPRA (test results compared to LLNA data)

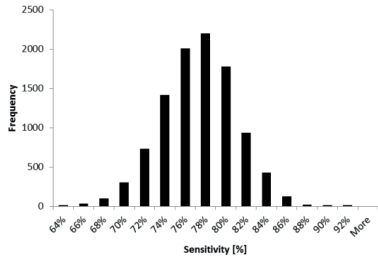


Figure C1: Distribution of sensitivity, derived from randomised samples including borderline substances

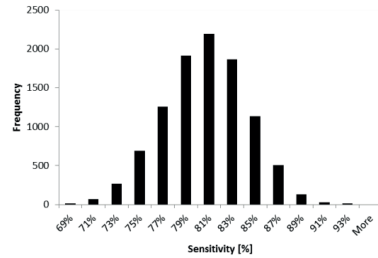


Figure C2: Distribution of sensitivity, derived from randomised samples excluding borderline substances

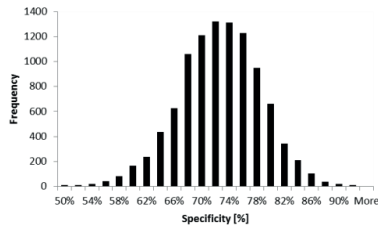


Figure C3: Distribution of specificity, derived from randomised samples including borderline substances

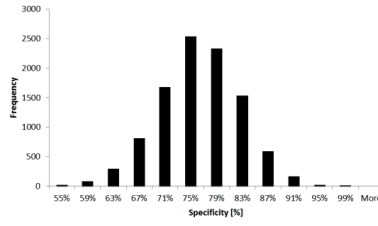


Figure C4: Distribution of specificity, derived from randomised samples excluding borderline substances

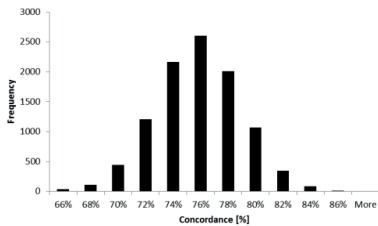


Figure C5: Distribution of concordance, derived from randomised samples including borderline substances

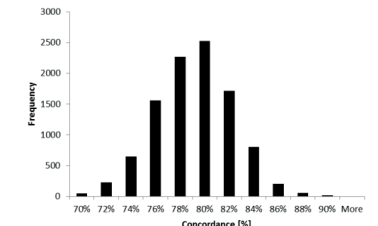


Figure C6: Distribution of concordance, derived from randomised samples excluding borderline substances

LuSens (using the LLNA as reference test)

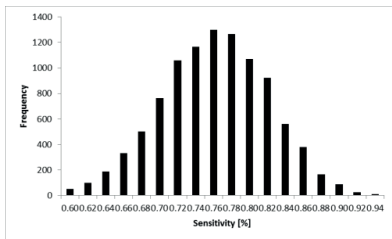


Figure C7: Distribution of sensitivity, derived

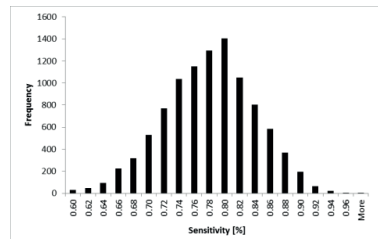


Figure C8: Distribution of the sensitivity,

from randomised samples including borderline substances

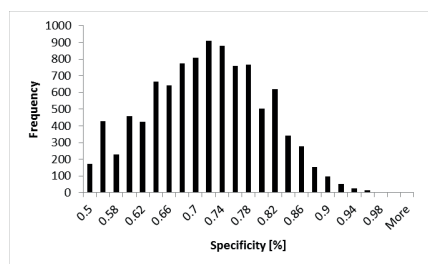


Figure C9: Distribution of specificity, derived from randomised samples including borderline substances

derived from randomised samples excluding borderline substances

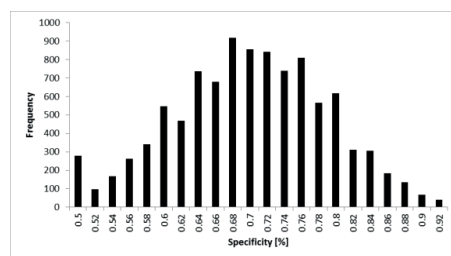


Figure C10: Distribution of the specificity, derived from randomised samples excluding borderline substances

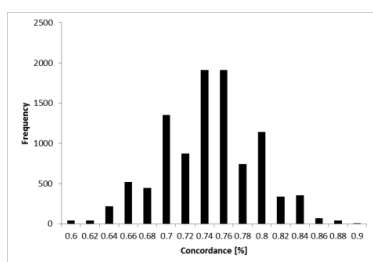


Figure C11: Distribution of the concordance, derived from randomised samples including borderline substances

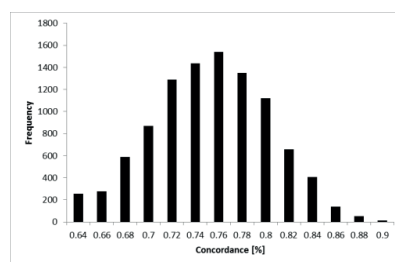


Figure C12: Distribution of the concordance, derived from randomised samples excluding borderline substances

h-CLAT (using the LLNA as reference test)

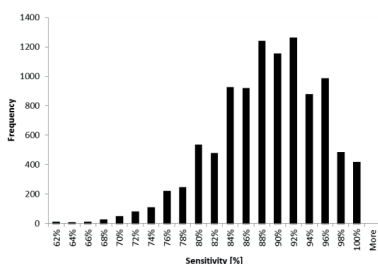


Figure C13: Distribution of the sensitivity, derived from randomised samples including borderline substances

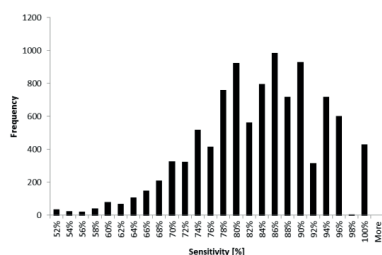


Figure C14: Distribution of the sensitivity, derived from randomised samples excluding borderline substances

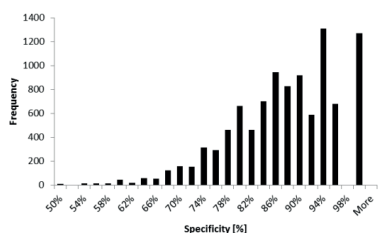


Figure C15: Distribution of the specificity,

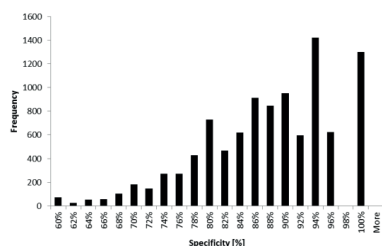


Figure C16: Distribution of the specificity,

derived from randomised samples including
borderline substances

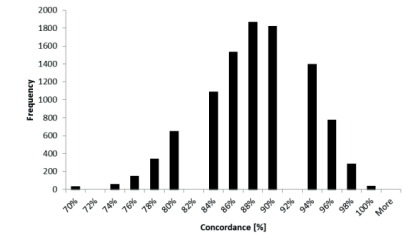


Figure C17: Distribution of the concordance,
derived from randomised samples including
borderline substances

derived from randomised samples excluding
borderline substances

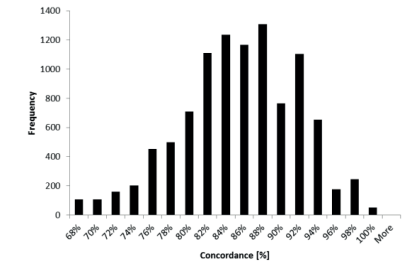


Figure C18: Distribution of the concordance,
derived from randomised samples excluding
borderline substances

“2 out of 3” ITS (using the LLNA as reference test)

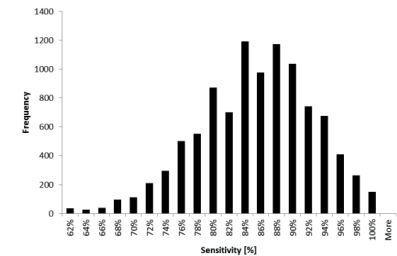


Figure C19: Distribution of the sensitivity,
derived from randomised samples including
borderline substances

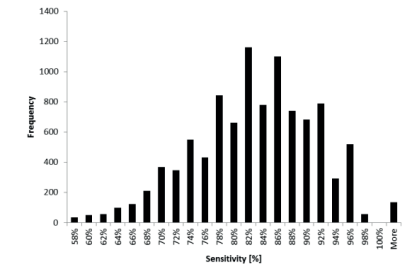


Figure C20: Distribution of the sensitivity,
derived from randomised samples excluding
borderline substances

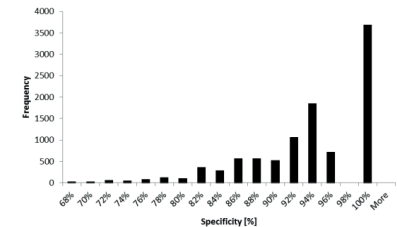


Figure C21: Distribution of the specificity,
derived from randomised samples including
borderline substances

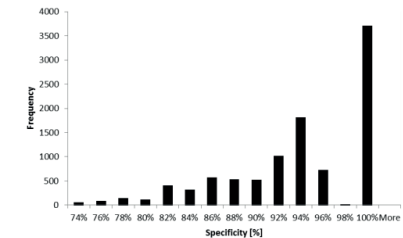


Figure C22: Distribution of the specificity,
derived from randomised samples excluding
borderline substance

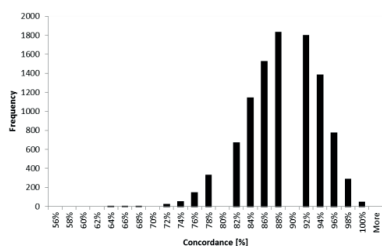


Figure C23: Distribution of the concordance, derived from randomised samples including borderline substances

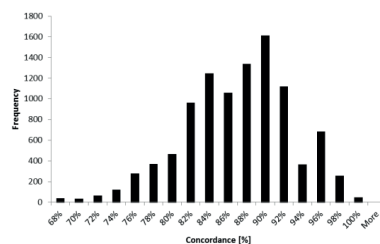


Figure C24: Distribution of the concordance, derived from randomised samples excluding borderline substances

Distributions of accuracy metrics for non-animal testing methods and the “2 out of 3”ITS - Using human data as reference test DPRA (using human data as reference test results)

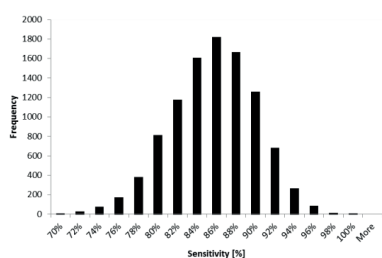


Figure C25: Distribution of sensitivity, derived from randomised samples including borderline substances

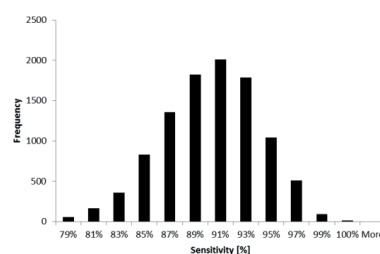


Figure C26: Distribution of sensitivity, derived from randomised samples excluding borderline substances

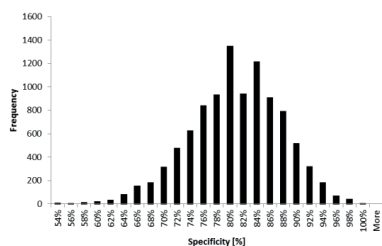


Figure C27: Distribution of specificity, derived from randomised samples including borderline substances

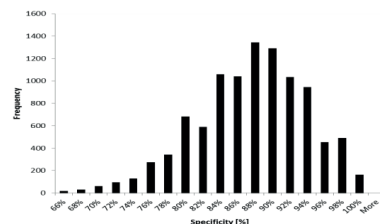


Figure C28: Distribution of specificity, derived from randomised samples excluding borderline substances

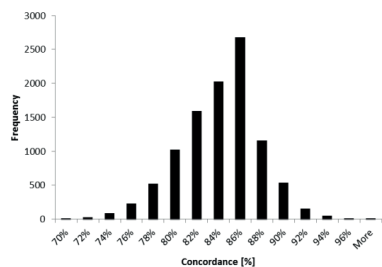


Figure C29: Distribution of concordance,

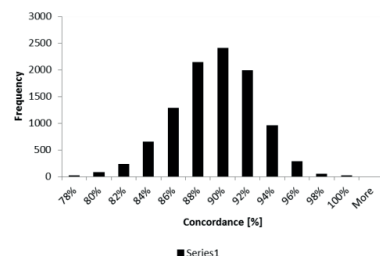


Figure C30: Distribution of concordance,

derived from randomised samples including
borderline substances

LuSens (using human data as reference test results)

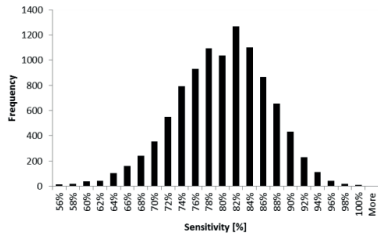


Figure C31: Distribution of sensitivity6,
derived from randomised samples including
borderline substances

derived from randomised samples excluding
borderline6 substances

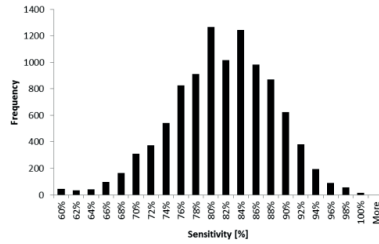


Figure C32: Distribution of 66the sensitivity,
derived from randomised samples excluding
borderline substances

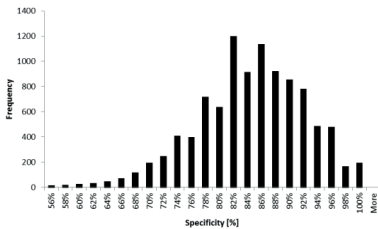


Figure C33: Distribution of specificity, derived
from randomised samples including borderline
substances

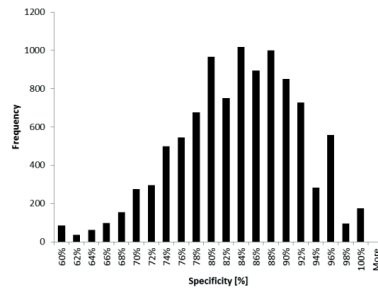


Figure C34: Distribution of the specificity,
derived from randomised samples excluding
borderline substances

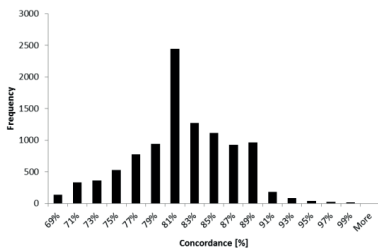


Figure C35: Distribution of the concordance,
derived from randomised samples including
borderline substances

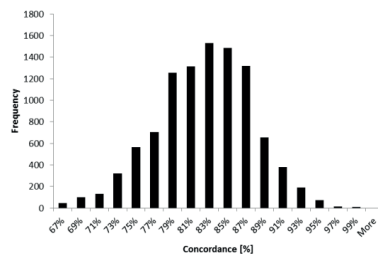


Figure C36: Distribution of the concordance,
derived from randomised samples excluding
borderline substances

h-CLAT (using human data as reference test results)

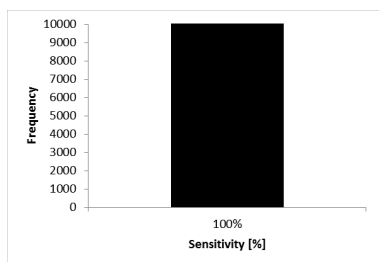


Figure C37: Distribution of the sensitivity, derived from randomised samples including borderline substances

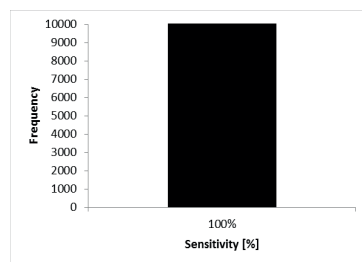


Figure C38: Distribution of the sensitivity, derived from randomised samples excluding borderline substances

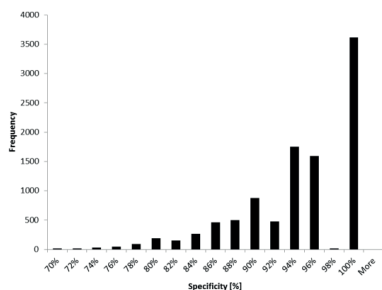


Figure C39: Distribution of the specificity, derived from randomised samples including borderline substances

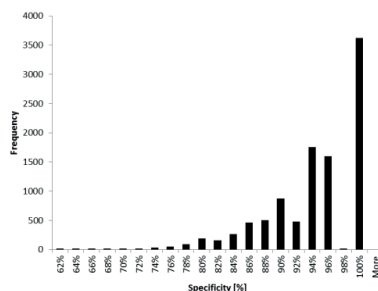


Figure C40: Distribution of the specificity, derived from randomised samples excluding borderline substances

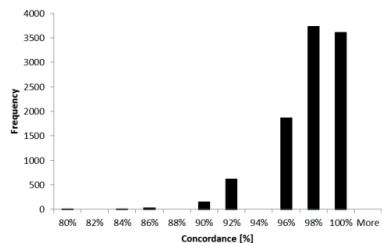


Figure C41: Distribution of the concordance, derived from randomised samples including borderline substances

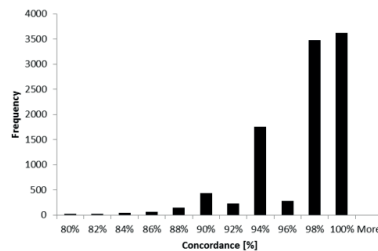


Figure C42: Distribution of the concordance, derived from randomised samples excluding borderline substances

"2 out of 3" ITS (using human data as reference test results)

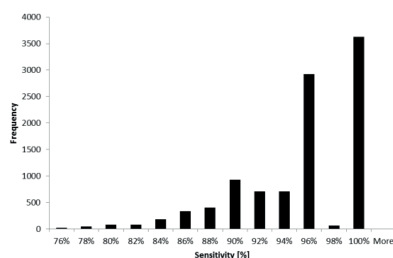


Figure C43: Distribution of the sensitivity, derived from randomised samples including

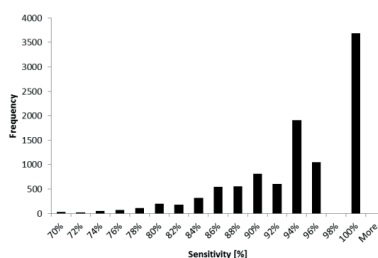


Figure C44: Distribution of the sensitivity, derived from randomised samples excluding

borderline substances

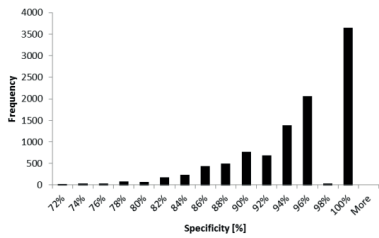


Figure C45: Distribution of the specificity, derived from randomised samples including borderline substances

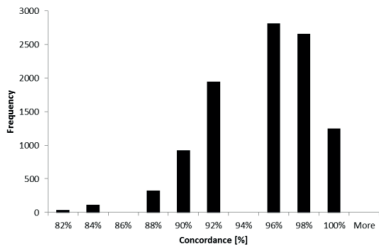


Figure C47: Distribution of the concordance, derived from randomised samples including borderline substances

borderline substances

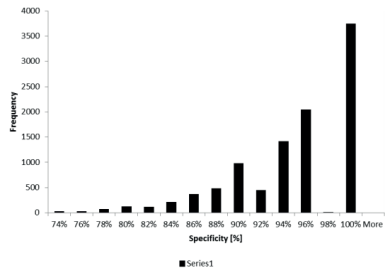
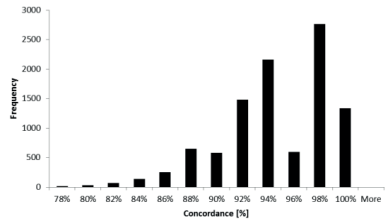


Figure C46: Distribution of the specificity, derived from randomised samples excluding borderline substances



FigureC48: Distribution of the concordance, derived from randomised samples excluding borderline substances

Appendix D

Accuracy metrics of testing methods when compared to human reference data

Table D1: Number of substances in the experimental samples (n) used for calculating accuracy metrics of non-animal testing methods and the “2 out of 3”ITS (test results compared to human data)

| | Including borderline substances | Excluding borderline substance |
|-----------------|---------------------------------|--------------------------------|
| DPRA | 107 | 95 |
| LuSens | 62 | 58 |
| h-CLAT | 35 | 28 |
| “2 out of 3”ITS | 35 | 32 |

Table D2: Minimum and maximum number of substances in randomised samples resulting from bootstrap resampling, after borderline substances were excluded (test results compared to human data)

| Randomised sample size (n) | DPRA | LuSens | h-CLAT | “2 out of 3” ITS |
|----------------------------|------|--------|--------|------------------|
| Min | 80 | 52 | 19 | 24 |
| Max | 105 | 62 | 35 | 36 |

Source: Own calculations.

Table D3: Accuracy metrics of the DPRA, LuSens, the h-CLAT and the “2 out of 3”ITS approach derived from experimental test results and randomised samples using bootstrap resampling (test results compared to human data)

| | | Experimental samples of substances | Randomised samples from bootstrap resampling | |
|--|-----------------|--|---|-----------|
| | | | Mean \pm SD | 95%CI |
| DPRA | | <i>n</i> =107 | | |
| | Sensitivity [%] | 85 | 85 \pm 4 | (76;93) |
| | Specificity [%] | 80 | 80 \pm 7 | (66;93) |
| | Concordance [%] | 83 | 83 \pm 4 | (76;90) |
| | | <i>n</i> =95 | | |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 89 | 89 \pm 4 | (81;96) |
| | Specificity [%] | 87 | 87 \pm 6 | (73;97) |
| | Concordance [%] | 88 | 88 \pm 4 | (82;95) |
| LuSens | | <i>n</i> =62 | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 79 | 79 \pm 7 | (65;91) |
| | Specificity [%] | 83 | 83 \pm 8 | (67;96) |
| | Concordance [%] | 81 | 81 \pm 5 | (69;90) |
| | | <i>n</i> =58 | | |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 80 | 80 \pm 7 | (66;92) |
| | Specificity [%] | 82 | 83 \pm 8 | (65;96) |
| | Concordance [%] | 81 | 81 \pm 5 | (71;91) |
| h-CLAT | | <i>n</i> =35 | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 100 | 100 \pm 0 | (100;100) |
| | Specificity [%] | 94 | 94 \pm 6 | (80;100) |
| | Concordance [%] | 97 | 97 \pm 3 | (91;100) |
| | | <i>n</i> =28 | | |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 100 | 100 \pm 0 | (100;100) |
| | Specificity [%] | 94 | 94 \pm 6 | (80;100) |
| | Concordance [%] | 96 | 96 \pm 4 | (88;100) |
| “2 out of 3” ITS | | <i>n</i> =35 | | |
| Complete samples (including borderline substances) | Sensitivity [%] | 95 | 95 \pm 5 | (82;100) |
| | Specificity [%] | 94 | 94 \pm 8 | (80;100) |
| | Concordance [%] | 94 | 94 \pm 4 | (86;100) |
| | | <i>n</i> =32 | | |
| Reduced sample (excluding borderline substances) | Sensitivity [%] | 93 | 93 \pm 7 | (78;100) |
| | Specificity [%] | 94 | 94 \pm 6 | (80;100) |
| | Concordance [%] | 94 | 94 \pm 4 | (84;100) |

Source: Own calculations.

Summary

Nowadays large amounts of chemicals are in use worldwide. Still, detailed knowledge about the hazardous properties has become available for just few chemicals. This has stimulated discussions among scientists and policy-makers on how adequate information about the hazards and risks of chemicals can be provided within a realistic time-frame while reducing testing costs and avoiding animal testing. During the past decade, much progress has been made in the development of non-animal testing methods for several toxicological endpoints. None of the existing non-animal testing methods, however, is considered to provide sufficient and adequate information to fully replace an animal test if applied as a standalone method. Using information from different non-animal methods integrated in forms of testing strategies, also called “Defined Approaches” (DAs), has been proposed as a promising solution to replace animal testing. Moreover, DAs have been widely considered to allow for an economically efficient way to assess hazards and risks, because they are assumed to deliver information about substances’ hazardous properties faster and cheaper than the “gold standard” animal tests, and to avoid the use of laboratory animals.

Efforts to develop non-animal testing strategies have given specific attention to the assessment of skin sensitisation which is the toxicological endpoint assessing a substance’s potential to cause acute contact dermatitis (ACD) in humans. This can be explained by several reasons. First, within the European Union, the REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) legislation prescribes skin sensitisation testing for industrial chemicals produced or imported in amounts more than one tonne per year. Second, substances used as ingredients in cosmetic products require skin sensitisation testing as a default. The Cosmetics Regulation, which entered into force in 2009, has enforced a marketing ban on all cosmetic products that contains chemicals tested in animal tests. This stimulated a scientific movement to develop non-animal methods and strategies for skin sensitisation assessment, based on *in vitro*, *in silico* and *in chemico* methods. Third, the adverse outcome pathway (AOP) which describes the sequence of biological and mechanistic events leading to the adverse outcome (i.e. the manifestation of ACD) is well-explored for skin sensitisation. The AOP has, therefore, been suggested as a guiding tool for the construction of testing strategies covering some or all key event in the skin sensitisation AOP. Still, there has been a need to gain better understanding in the conceptual requirements to develop resource-efficient testing strategies from an economic perspective. In this context, key conceptual challenges focus on (i) which non-animal testing methods to select in order to provide adequate hazard information, (ii) how to combine these methods into a testing strategy, and (iii) when testing

should stop. This also requires that non-animal testing methods' precision, and uncertainties underlying to their predictivity, need to be well- understood and transparently documented.

The objective of this thesis is to develop and apply an economic approach to the development of optimised non-animal testing strategies used for skin sensitisation potential (hazard) assessment. Furthermore, the thesis examines the uncertainty of test information due to the biological and technical variability and explores the impact of this uncertainty on non-animal testing methods predictive accuracy. Predictive accuracy is the ability of a non-animal testing method to predict the test result of a reference animal test for tested substances.

To provide a systematic overview of the current state-of-work regarding non-animal testing for skin sensitisation, Chapter 2 surveys the criteria suggested in the toxicological literature for the development of testing strategies. Furthermore, Chapter 2 suggests conceptual criteria and informational requirements in order to develop resource-efficient testing strategies. As a follow-up step existing testing strategies (i.e. DAs) combining information from different (non-animal) testing methods for skin sensitisation potential or potency, are qualitatively evaluated applying conceptual and informational criteria. We observe that existing testing strategies for skin sensitisation assessment focus predominantly on the maximisation of information, which is related to mechanistic criteria e.g. either covering the key events in the AOP or increasing the predictive accuracy estimates. Although the need to account for direct and indirect costs of testing has been widely acknowledged in the literature, cost components have been largely ignored in the development of integrated testing strategies. Optimising toxicity testing requires, however, balancing information outcomes with costs. The challenge is, therefore, to develop a methodological approach that guides the construction of resource efficient non-animal testing strategies.

Chapter 3 introduces a Bayesian Value of Information (VOI) model as an economic approach to the optimisation of non-animal testing strategies for the assessment of skin sensitisation potential. A set of non-animal testing methods (i.e. the DPRA, LuSens, KeratinoSens™, the h-CLAT and the OECD toolbox) are evaluated according to their Expected Value of Test Information (EVTI), which quantifies the expected net welfare gains from decision-making upon the use of a substance with additional information from testing. The EVTI is quantified for all individual non-animal testing methods, and their combinations into battery- and sequential combinations. Results are compared to those of the reference animal test (LLNA). Expected welfare gains or losses from using a (toxic) substance substances are approximated by estimating health damage costs from ACD caused, using the case of cosmetic

ingredients (i.e. Kathon CG) as an illustrative example. The Bayesian VOI model offers a probabilistic method guiding (i) the selection of testing methods, (ii) the order of testing methods and (iii) when testing should stop. The stopping rule is an endogenous component of the VOI model. Testing has a positive (economic) value if and only if the information gains net of testing costs are positive. Chapter 3 concludes that combinations of non-animal testing methods into either batteries or sequential strategies reveal a higher EVTI than the reference animal test. Furthermore, it can be shown that resource efficient testing strategies do not need to cover all key events and do not have to follow the order of key events in the skin sensitisation AOP. Rather, the optimal selection of testing methods depends on the interplay of multiple parameters, such as a decision-maker's prior beliefs upon the true properties of a substance, the predictive accuracy of testing methods, the expected welfare gains from marketing a substance, and, finally, testing costs.

Chapter 4 examines the impact of biological and technical variability on the precision of testing methods for assessing skin sensitisation potential. Precision denotes a non-animal testing method's ability to show concordant results in repeated applications. In general, for classification purposes conclusions on the hazardous properties of chemicals are based on binary "positive/negative" outcomes. Binary outcomes are derived by applying classification thresholds into continuous readouts from testing. However, for substances for which test results are close to the classification threshold non-animal testing methods, but also animal tests, can deliver discordant results in repeated testing. More specifically, we can quantify a range to the left and the right of the classification threshold within which discordant results can be expected with a certain probability. This range is called "grey zone" or "borderline range" around the classification thresholds of testing methods' prediction models. In Chapter 4 we quantify the borderline range for the LLNA as the reference animal test, the DPRA, LuSens, and the h-CLAT as non-animal testing methods, and for a combination of these methods into an integrated testing strategy i.e. the "2 out of 3" ITS. Furthermore, we identify the number of substances in the experimental samples of these methods for which test outcomes fall within the borderline range. Chapter 4 concludes that the technical and biological variability of testing methods impacts the precision of testing methods. Substances which are borderline, i.e. which revealed test results in the borderline range of the testing method, cannot unambiguously be classified as "hazardous/non-hazardous" thus indicating its ability to cause an effect or not. For such substances, a clear-cut classification is, therefore, not possible. Rather, further testing might be required to gain additional evidence on a substance's intrinsic properties. Chapter 4 suggests that the borderline range should be

quantified and documented as a default for non-animal testing methods and the animal tests to provide transparent information about testing methods' precision.

Chapter 5 examines the impact of different types of uncertainties in testing methods' predictive accuracy. In particular, the impact of limited precision, and the impact of varying sample size and composition is examined in three steps. First, we examine the impact of non-animal testing methods' limited precision on predictive accuracy metrics. This is done by comparing sensitivity, specificity and concordance derived from experimental substances' samples including borderline substances (i.e. the "complete" samples) with accuracy metrics derived after excluding borderline substances (i.e. the "reduced" samples). Second, we examine the impact of sample composition on accuracy metrics. Comparing accuracy metrics from randomised "complete" and "reduced" samples capture the joint effect from sample composition and limited precision. Third, we examine the joint impact of limited precision, variations in sample composition and variations of sample size on non-animal testing methods' accuracy. To create randomised samples we use the non-parametric bootstrap analysis. The analysis is applied to experimental samples tested with the DPRA, LuSens, the h-CLAT and the "2 out of 3" ITS. Results suggest that the impact of limited precision, sample size and composition on non-animal testing methods' predictive accuracy depends on the relationship of borderline substances and substances with a clear-cut classification (non-borderline) in experimental samples. Chapter 5 suggests using ranges for the accuracy metrics, rather than point estimates, to better reflect uncertainties, and to facilitate a transparent comparison, of non-animal methods' predictive accuracy in a regulatory context.

Using skin sensitisation as an illustrative example, this research has shown the importance of applying an economic approach to the development of testing strategies. Furthermore, this thesis has offered novel insights regarding the impact of different types of uncertainty on non-animal testing methods' predictive accuracy, which is a key parameter for determining the information outcomes and, thus, the "informational value" from testing. These aspects are relevant for the evaluation of individual testing methods and for guiding the optimisation of non-animal testing strategies, for both scientific and regulatory purposes. The insights offered in this thesis, therefore, support the development of optimised, i.e. resource efficient, approaches to toxicological testing ensuring better-informed decision-making for a safe use of chemicals.

Acknowledgements

I would like to extend my thanks to everyone who was involved in this PhD project. First of all, thanks go to my promotor Ekko van Ierland and my co-promotor Silke Gabbert. Many thanks to both of you for the guidance towards the completion of my thesis and the careful revisions of our papers. I really appreciate the patience you showed in times where I was struggling with health issues.

I extend my thanks to my co-promotor Robert Landsiedel who trusted me with this project. Thanks go to my colleagues Susanne Kolle, Daniel Urbisch, Annette Mehling, Dennis Mulliner, and Tzutzuy Ramirez with whom I had such a good collaboration while I was staying at the Experimental Toxicology and Ecology Department at BASF.

I would like to thank my supervisor Andrew Worth, who provided me with valuable comments on my research and supported me in difficult times during my visit at the Institute for Health and Consumer Protection (IHCP) of the Joint Research Centre (JRC) in Italy. During my stay at the JRC I also worked with nice people who helped me further my research. Silvia Casati and Roman Liska provided fruitful comments in parts of my research.

Thanks go to my colleagues in the Environmental Economics and Natural Resources (ENR) Group. Wil and Gré thank you for the administrative support. I would also like to acknowledge the senior people of ENR. Our discussions during coffee breaks, triggered me to improve my research. Hans-Peter Weikard provided me with valuable comments to improve the propositions belonging to this thesis. Jeroen Klomp also provided me with interesting comments with regard to the methodology followed in parts of my research.

Special thanks go to my fellow PhD students. Coffee and lunch breaks have been a source of courage and strength to keep on going till the end. I know, you expect to read your names at this point, but I will leave it up to your judgment. If reading this text enables a smile on your face and brings some nice memories to you, then you know I was thinking of you while typing these words. Admittedly, not the most sentimental piece of text but, really, you should know that you added joyful moments all those years, and I feel grateful for that.

Όσο για τους φίλους, αυτούς τους ξεχωριστούς που είναι εδώ ακόμα και μοιράστηκαν τα καλά και τα κακά των τελευταίων χρόνων, σας στέλνω την αγαπη μου καθημερινά, έστω και από απόσταση κατά περιόδους. Τελευταία και πλέον σημαντική, αφήσα να ευχαριστήσω την οικογένεια μου. Η οικογένεια μου με έμαθε να μη φοβάμαι, να υπερασπιζομαι τις επιλογές μου, και να σεβομαι τις ευθύνες που αναλαμβάνω.

About the author

Maria Leontaridou grew up on Ikaria Island, Greece, where she finished her high school education. She succeeded in the University of the Aegean where she did her bachelor studies on environmental sciences. During her bachelor studies she got involved in several projects as an undergraduate research assistant and she graduated in 2009. After that, she continued with her master studies at Wageningen University, focusing on topics of toxicology, where she graduated in 2011. Shortly after the completion of her master studies, Maria worked as a trainee at the Emerging Risks Unit (EmRisk) of the European Food Safety Authority (EFSA) of the European Union (EU). In 2012 she received the responsibility to undergo her PhD research in collaboration with the Environmental Economics Group (ENR) of Wageningen University, the Institute for Health and Consumer Protection (IHCP) of the Joint Research Centre (JRC) of the EU and the Department of Experimental Toxicology and Ecology of BASF SE. During her PhD, Maria mainly stayed at the ENR in Wageningen while she also spent several months at both BASF and the JRC.

References

- Adler, S., D. Basketter, S. Creton, et al. (2011). "Alternative (non-animal) methods for cosmetics testing: Current status and future prospects-2010." *Archives of Toxicology* 85(5): 367-485.
- Agnese, G., D. Risso and S. De Flora (1984). "Statistical evaluation of inter- and intra-laboratory variations of the Ames test, as related to the genetic stability of *Salmonella* tester strains." *Mutat Res* 130(1): 27-44.
- Ahlers, J., F. Stock and B. Werschkun (2008). "Integrated testing and intelligent assessment—new challenges under REACH." *Environmental Science and Pollution Research* 15(7): 565-572.
- Andersen, M. E. and D. Krewski (2010). "The Vision of Toxicity Testing in the 21st Century: Moving from Discussion to Action." *Toxicological Sciences* 117(1): 17-24.
- Ankley, G. T., R. S. Bennett, R. J. Erickson, et al. (2010). "Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment." *Environmental Toxicology and Chemistry* 29(3): 730-741.
- Ashikaga, T., Y. Yoshida, M. Hirota, et al. (2006). "Development of an in vitro skin sensitization test using human cell lines: The human Cell Line Activation Test (h-CLAT): I. Optimization of the h-CLAT protocol." *Toxicology in Vitro* 20(5): 767-773.
- Ashikaga, T., H. Sakaguchi, S. Sono, 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 Alternatives to Laboratory Animals* 38(4): 275-284.
- Asturiol, D., S. Casati and A. Worth (2016). "Consensus of classification trees for skin sensitisation hazard prediction." *Toxicology in Vitro* 36: 197-209.
- Augustin, M. and I. Zschocke (2001). "Quality of life and economy of allergic skin diseases." *Allergologie* 24(9): 433-442.
- Balls, M., P. Amcoff, S. Bremer, et al. (2006). "The Principles of Weight of Evidence Validation of Test Methods and Testing Strategies: The Report and Recommendations of ECVAM Workshop 58." *Alternatives to laboratory animals : ATLA* 34(6): 603-620.
- Balls, M., R. D. Combes and N. Bhogal (2012). "The use of integrated and intelligent testing strategies in the prediction of toxic hazard and in risk assessment." *Advances in Experimental Medicine and Biology* 745: 221-253.
- Basketter, D., J. Crozier, B. Hubesch, et al. (2012). "Optimised testing strategies for skin sensitization – The LLNA and beyond." *Regulatory Toxicology and Pharmacology* 64(1): 9-16.
- Basketter, D., N. Alepee, S. Casati, et al. (2013). "Skin sensitisation--moving forward with non-animal testing strategies for regulatory purposes in the EU." *Regul Toxicol Pharmacol* 67(3): 531-535.
- Basketter, D., A., N. Alepee, T. Ashikaga, et al. (2014). "Categorization of chemicals according to their relative human skin sensitizing potency." *Dermatitis* 25(1): 11-21.
- Bauch, C., S. N. Kolle, T. Ramirez, et al. (2012). "Putting the parts together: combining in vitro methods to test for skin sensitizing potentials." *Regul Toxicol Pharmacol* 63(3): 489-504.
- Bergstrom, T. C. and H. R. Varian (2003). *Intermediate Microeconomics: Instructor's Manual*, Norton.
- Bjørner, J. and H. Keiding (2004). "Cost-effectiveness with multiple outcomes." *Health Economics* 13(12): 1181-1190.
- Boobis, A. R., J. E. Doe, B. Heinrich-Hirsch, et al. (2008). "IPCS framework for analyzing the relevance of a noncancer mode of action for humans." *Crit Rev Toxicol* 38(2): 87-96.
- Both, H., M. L. Essink-Bot, J. Busschbach, et al. (2007). "Critical Review of Generic and Dermatology-Specific Health-Related Quality of Life Instruments." *Journal of Investigative Dermatology* 127(12): 2726-2739.
- Bottini, A. A. and T. Hartung (2009). "Food for thought... on the economics of animal testing." *Altex* 26(1): 3-16.
- Casati, S., W. A. Amcoff P., et al. (2013). *EURL ECVAM Strategy for Replacement of Animal Testing for Skin Sensitisation Hazard Identification and Classification EUR – Scientific and Technical Research series*. Luxembourg: Publications Office of the European Union - ISSN 1018-5593 (print), ISSN 1831-9424 (online).
- CEFIC (2015). *Joint cross-sector workshop on alternatives for skin sensitization testing and assessment*. Workshop report. Helsinki / Finland.
- Claxton, K. (1999). "Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals." *Health Economics* 8(3): 269-274.
- Claxton, K., L. Ginnelly, M. Sculpher, et al. (2004). "A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme." *Health Technology Assessment* 8(31): iii-60.
- Clemen, R. T. and T. Reilly (2001). *"Making Hard Decisions with Decision Tools"*. Belmont CA: Duxbury Press.
- Cooper, J. A., R. Saracci and P. Cole (1979). "Describing the validity of carcinogen screening tests." *British Journal Of Cancer* 39(1): 87-89.
- Cunningham, S. J. (2001). "An Introduction to Economic Evaluation of Health Care." *Journal of Orthodontics* 28(3): 246-250.
- Daston, G., D. J. Knight, M. Schwarz, et al. (2015). "SEURAT: Safety Evaluation Ultimately Replacing Animal Testing—Recommendations for future research in the field of predictive toxicology." *Archives of Toxicology* 89(1): 15-23.
- De Wever, B., H. W. Fuchs, M. Gaca, et al. (2012). "Implementation challenges for designing integrated in vitro testing strategies (ITS) aiming at reducing and replacing animal experimentation." *Toxicol In Vitro* 26(3): 526-534.
- Dellarco, V., T. Henry, P. Sayre, et al. (2010). "Meeting the common needs of a more effective and efficient testing and assessment paradigm for chemical risk management." *J Toxicol Environ Health B Crit Rev* 13(2-4): 347-360.
- Dellarco, V. and P. A. Fenner-Crisp (2012). "Mode of action: moving toward a more relevant and efficient assessment paradigm." *J Nutr* 142(12): 2192s-2198s.
- Dimitrov, S., A. Detroyer, C. Piroird, et al. (2016). "Accounting for data variability, a key factor in in vivo/in vitro relationships: application to the skin sensitization potency (in vivo LLNA versus in vitro DPRA) example." *Journal of Applied Toxicology*: n/a-n/a.
- Dimitrov, S., D., L. K. Low, G. Y. Patlewicz, et al. (2005). "Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates." *Int J Toxicol* 24(4): 189-204.

References

- Dumont, C., J. Barroso, I. Matys, et al. (2016). "Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches." *Toxicology in Vitro* 34: 220-228.
- EC (2003a). "Directive 2003/15/EC of the European Parliament and of the council of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products."
- EC (2003b). "DIRECTIVE 2010/63/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 September 2010 on the protection of animals used for scientific purposes."
- EC (2006). "Regulation (EC) No 1907/2006 of the European Parliament and the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC."
- EC (2008). "REGULATION (EC) No 1272/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006."
- EC (2009). "Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products."
- EC (2016). "Commission Regulation (EU) 2016/1179 of 19 July 2016 amending, for the purposes of its adaptation to technical and scientific progress, Regulation (EC) No 1272/2008 of the European Parliament and of the Council on classification, labelling and packaging of substances and mixtures (Text with EEA relevance)."
- ECHA (2014a). "The Use of Alternatives to Testing on Animals for the REACH Regulation. Second report under Article 117(3) of the REACH Regulation." European Chemicals Agency, Helsinki.
- ECHA (2014b). "Stated-preference study to examine the economic value of benefits of avoiding selected adverse human health outcomes due to exposure to chemicals in the European Union; Part I: sensitization & dose toxicity." Service contract for the European Chemicals Agency No. ECHA/2011/123.
- EC - Eurostat. "Eurostat". Available at: <http://ec.europa.eu/eurostat> (Accessed 20.06.16).
- ECHA (2016). "Guidance on information requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint specific guidance Draft Version 5.0."
- ECVAM (2012). "Direct Peptide Reactivity Assay (DPRA) ECVAM Validation Study Report."
- ECVAM (2013). "Human Cell Line Activation Test (h-CLAT) Validation Study Report."
- Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. (Monographs on Statistics and Applied Probability). London, UK, Chapman and Hall/CRC.
- Ellison, C. M., J. C. Madden, P. Judson, et al. (2010). "Using In Silico Tools in a Weight of Evidence Approach to Aid Toxicological Assessment." *Molecular Informatics* 29(1-2): 97-110.
- Emter, R., G. Ellis and A. Natsch (2010). "Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro." *Toxicology and Applied Pharmacology* 245(3): 281-290.
- Ennever, F. K., H. S. Rosenkranz, L. B. Lave, et al. (1990). "Value-of-information analysis of testing strategies: estimating the effect of uncertainty about the proportion of chemicals that are true human carcinogens." *Progress in Clinical and Biological Research* 340D: 295-304.
- EPA (2005). *Guidelines for Carcinogen Risk Assessment*. Risk Assessment Forum Washington, DC, U.S. Environmental Protection Agency
- Gabbert, S. and H. P. Weikard (2010). "A theory of chemicals regulation and testing." *Natural Resources Forum* 34(2): 155-164.
- Gabbert, S. and E. C. van Ierland (2010). "Cost-effectiveness analysis of chemical testing for decision-support: How to include animal welfare?" *Human and Ecological Risk Assessment* 16(3): 603-620.
- Gabbert, S. and H. P. Weikard (2013). "Sequential Testing of Chemicals when Costs Matter: A Value of Information Approach." *Human and Ecological Risk Assessment* 19(4): 1067-1088.
- Gerberick, G. F., J. D. Vassallo, R. E. Bailey, et al. (2004). "Development of a Peptide Reactivity Assay for Screening Contact Allergens." *Toxicological Sciences* 81(2): 332-334.
- Gerberick, G. F., J. D. Vassallo, L. M. Foertsch, et al. (2007). "Quantification of Chemical Peptide Reactivity for Screening Contact Allergens: A Classification Tree Model Approach." *Toxicological Sciences* 97(2): 417-427.
- Gocht, T., E. Berggren, H. J. Ahr, et al. (2015). "The SEURAT-1 approach towards animal free human safety assessment." *Altex* 32(1): 9-24.
- Goebel, C., P. Aeby, N. Ade, et al. (2012). "Guiding principles for the implementation of non-animal safety assessment approaches for cosmetics: Skin sensitisation." *Regulatory Toxicology and Pharmacology* 63(1): 40-52.
- Gomes, C., Noçairi, H., Thomas, M., Collin, J.F., Ibanez, F., Saporta, G. (2012). "Stacking prediction for a binary outcome." *COMPSTAT, 20th International Conference on Computational Statistics*, Limassol, 271-282.
- Grandjean, P. (2015). "Toxicology research for precautionary decision-making and the role of Human & Experimental Toxicology." *Human and Experimental Toxicology* 34(12): 1231-1237.
- Grindon, C., R. Combes, M. T. Cronin, et al. (2006a). "Integrated testing strategies for use in the EU REACH system." *Alternatives to laboratory animals : ATLA* 34(4): 407-427.
- Grindon, C., R. Combes, M. T. D. Cronin, et al. (2006b). "Integrated testing strategies for use in the EU REACH system." *Alternatives to laboratory animals : ATLA* 34(4): 407-427.
- Grindon, C., R. Combes, M. T. Cronin, et al. (2008a). "Integrated decision-tree testing strategies for developmental and reproductive toxicity with respect to the requirements of the EU REACH legislation." *Alternatives to laboratory animals : ATLA* 36 Suppl 1: 123-138.
- Grindon, C., R. Combes, M. T. Cronin, et al. (2008b). "An integrated decision-tree testing strategy for eye irritation with respect to the requirements of the EU REACH legislation." *Alternatives to laboratory animals : ATLA* 36 Suppl 1: 111-122.

- Grindon, C., R. Combes, M. T. Cronin, et al. (2008c). "Integrated decision-tree testing strategies for skin corrosion and irritation with respect to the requirements of the EU REACH legislation." *Alternatives to laboratory animals : ATLA* 36 Suppl 1: 65-74.
- Grindon, C., R. Combes, M. T. D. Cronin, et al. (2008d). "An integrated decision-tree testing strategy for repeat dose toxicity with respect to the requirements of the EU REACH legislation." *ATLA Alternatives to Laboratory Animals* 36(SUPPL. 1): 139-147.
- Grindon, C., R. Combes, M. T. Cronin, et al. (2008e). "An integrated decision-tree testing strategy for skin sensitisation with respect to the requirements of the EU REACH legislation." *Alternatives to laboratory animals : ATLA* 36 Suppl 1: 75-89.
- Groh, K. J., R. N. Carvalho, J. K. Chipman, et al. (2015). "Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: II. A focus on growth impairment in fish." *Chemosphere* 120(0): 778-792.
- Halvarsson, K. and M. Loden (2007). "Increasing quality of life by improving the quality of skin in patients with atopic dermatitis." *International Journal of Cosmetic Science* 29(2): 69-83.
- Hansen, B. and M. Blainey (2008). "Registration: The Cornerstone of REACH." *Review of European Community & International Environmental Law* 17(1): 107-125.
- Hansson, S. O. and C. Rudén (2007). "Towards a theory of tiered testing." *Regulatory Toxicology and Pharmacology* 48(1): 35-44.
- Hartung, T. (2010a). "Toxicology for the twenty-first century." *Nature* 460(7252): 208-212.
- Hartung, T. (2010b). "Comparative analysis of the revised Directive 2010/63/EU for the protection of laboratory animals with its predecessor 86/609/EEC - a t4 report." *Altex* 27(4): 285-303.
- Hartung, T., T. Luechtefeld, A. Maertens, et al. (2013). "Food for thought: Integrated testing strategies for safety assessments." *Altex* 30(1): 3-18.
- Held L. and D. S. Bové (2014). *Applied statistical inference. Likelihood and Bayes*. Berlin Heidelberg, Springer.
- Heringa, M. B., L. De Wit-Bos, P. Bos, et al. (2015). Do current EU regulations for the safety assessment of chemical substances pose legal barriers for the use of alternatives to animal testing? R. R. 2014-0148.
- Hirota, M., H. Kouzuki, T. Ashikaga, et al. (2013). "Artificial neural network analysis of data from multiple in vitro assays for prediction of skin sensitization potency of chemicals." *Toxicology in Vitro* 27(4): 1233-1246.
- Hirota, M., S. Fukui, K. Okamoto, et al. (2015). "Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization." *J Appl Toxicol* 35(11): 1333-1347.
- Hirshleifer, J. C. F. p. d. S. (1971). "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review* 61(4): 561-574.
- Hoffmann, S. and T. Hartung (2005). "Diagnosis: Toxic! - Trying to Apply Approaches of Clinical Diagnostics and Prevalence in Toxicology Considerations." *Toxicological Sciences* 85(1): 422-428.
- Hoffmann, S. and T. Hartung (2006). "Toward an evidence-based toxicology." *Human and Experimental Toxicology* 25(9): 497-513.
- Hoffmann, S., A. G. Saliner, G. Patlewicz, et al. (2008). "A feasibility study developing an integrated testing strategy assessing skin irritation potential of chemicals." *Toxicology Letters* 180(1): 9-20.
- Hoffmann, S. (2015). LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *Altex* 32: 379-383.
- Hongbo, Y., C. L. Thomas, M. A. Harrison, et al. (2005). "Translating the Science of Quality of Life into Practice: What Do Dermatology Life Quality Index Scores Mean[quest]." 125(4): 659-664.
- Hothorn, L. A. (2002). "Selected biostatistical aspects of the validation of in vitro toxicological assays." *Altern Lab Anim* 30 Suppl 2: 93-98.
- Hothorn, L. A. (2003). "Statistics of interlaboratory in vitro toxicological studies." *Altern Lab Anim* 31 Suppl 1: 43-63.
- Howson, C. and P. Urbach (1991). "Bayesian reasoning in science." *Nature* 350(6317): 371-374.
- Hristozov, D. R., A. Zabeo, C. Foran, et al. (2014). "A weight of evidence approach for hazard screening of engineered nanomaterials." *Nanotoxicology* 8(1): 72-87.
- Hurd, G. (2015). Planning a cost-effectiveness study. *Integrated Primary and Behavioral Care: Role in Medical Homes and Chronic Disease Management*: 115-136.
- Hurley, J., J. C. Anthony and P. N. Joseph (2000). Chapter 2 - An Overview of the Normative Economics of the Health Sector*. *Handbook of Health Economics*, Elsevier. Volume 1, Part A: 55-118.
- ICCVAM (2009). "Recommended Performance Standards: Murine Local Lymph Node Assay. NIH Publication Number 09-7357. Research Triangle Park, NC: National Institute of Environmental Health Sciences."
- Jackson, K. D., L. D. Howie and L. J. Akinbami (2013). "Trends in allergic conditions among children: United States, 1997-2011." *NCHS Data Brief*(121): 1-8.
- Jaworska, J. and S. Hoffmann (2010). "Integrated Testing Strategy (ITS) - Opportunities to better use existing data and guide future testing in toxicology." *Altex* 27(4): 231-242.
- Jaworska, J., S. Gabbert and T. Aldenberg (2010). "Towards optimization of chemical testing under REACH: A Bayesian network approach to Integrated Testing Strategies." *Regulatory Toxicology and Pharmacology* 57(2-3): 157-167.
- Jaworska, J., A. Harol, P. S. Kern, et al. (2011). "Integrating non-animal test information into an adaptive testing strategy - Skin sensitization proof of concept case." *Altex* 28(3): 211-225.
- Jaworska, J., Y. Dancik, P. Kern, et al. (2013). "Bayesian integrated testing strategy to assess skin sensitization potency: From theory to practice." *Journal of Applied Toxicology* 33(11): 1353-1364.
- Jaworska, J. (2016). "Integrated Testing Strategies for Skin Sensitization Hazard and Potency Assessment—State of the Art and Challenges." *Cosmetics* 3(2): 16.
- Jaworska, J. S., A. Natsch, C. Ryan, et al. (2015). "Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy." *Archives of Toxicology* 89(12): 2355-2383.

References

- Jones, A. M., J. C. Anthony and P. N. Joseph (2000). Chapter 6 - Health Econometrics*. Handbook of Health Economics, Elsevier. Volume 1, Part A: 265-344.
- Kimber, I., R. J. Dearman, E. W. Scholes, et al. (1994). "The local lymph node assay: developments and applications." *Toxicology* 93(1): 13-31.
- Kimber, I., D. A. Basketter, K. Berthold, et al. (2001). "Skin sensitization testing in potency and risk assessment." *Toxicological Sciences* 59(2): 198-208.
- Kimber, I., D. A. Basketter, G. F. Gerberick, et al. (2002). "Allergic contact dermatitis." *International Immunopharmacology* 2(2-3): 201-211.
- Kinsner-Ovaskainen, A., Z. Akkan, S. Casati, et al. (2009). "Overcoming barriers to validation of non-animal partial replacement methods/integrated testing strategies: The report of an EPAA-ECVAM workshop." *ATLA Alternatives to Laboratory Animals* 37(4): 437-444.
- Kleinstreuer, N. C., K. Sullivan, D. Allen, et al. (2016). "Adverse outcome pathways: From research to regulation scientific workshop report." *Regulatory Toxicology and Pharmacology* 76: 39-50.
- Koch, L. and N. A. Ashford (2006). "Rethinking the role of information in chemicals policy: implications for TSCA and REACH." *Journal of Cleaner Production* 14(1): 31-46.
- Kolle, S. N., D. A. Basketter, S. Casati, et al. (2013). "Performance standards and alternative assays: Practical insights from skin sensitization." *Regulatory Toxicology and Pharmacology* 65(2): 278-285.
- Kopp-Schneider, A., P. Prieto, A. Kinsner-Ovaskainen, et al. (2013). "Design of a testing strategy using non-animal based test methods: Lessons learnt from the ACuteTox project." *Toxicology in Vitro* 27(4): 1395-1401.
- Krewski, D., M. E. Andersen, E. Mantus, et al. (2009). "Toxicity testing in the 21st century: Implications for human health risk assessment: Perspective." *Risk Analysis* 29(4): 474-479.
- Krewski, D., D. Acosta, Jr., M. Andersen, et al. (2010). "Toxicity Testing in the 21st Century: A Vision and a Strategy." *Journal of Toxicology and Environmental Health, Part B* 13(2-4): 51-138.
- Krzyszowski W. L. and D. J. Hand. (2009). ROC curves for continuous data London, CRC Press Taylor & Francis Group
- Landesmann, B., M. Mennecozzi, E. Berggren, et al. (2013). "Adverse outcome pathway-based screening strategies for an animal-free safety assessment of chemicals." *Alternatives to laboratory animals : ATLA* 41(6): 461-471.
- Lave, L., Ennever FK, et al. (1988). "Information value of the rodent bioassay." *Nature* 336: (6200): 631-633.
- Leiva-Salinas, M., L. Frances, I. Marin-Cabanas, et al. (2014). "Methylchloroisothiazolinone/methylisothiazolinone and methylisothiazolinone allergies can be detected by 200 ppm of methylchloroisothiazolinone/ methylisothiazolinone patch test concentration." *Dermatitis : contact, atopic, occupational, drug* 25(3): 130-134.
- Leontaridou, M., D. Urbisch, S. N. Kolle, et al. (2017a). "The borderline range of toxicological methods: Quantification and implications for evaluating precision The borderline range of prediction models for skin sensitisation potential assessment: Quantification and implications for evaluating non-animal testing methods' precision" *ALTEX* in press doi: 10.14573/altex.1606271...
- Leontaridou, M., S. Gabbert, E. C. van Ierland, et al. (2016). "Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian Value-of-Information analysis." *Altern Lab Anim* 44(3): 255-269.
- Leontaridou, M., S. Gabbert and R. Landsiedel (2017b). "The uncertainties in measures of predictivity: The impact of sample size and precision of non-animal methods for skin sensitisation." In preparation. (manuscript is available from the author on request)
- Lewis, A., N. Kazantzis, I. Fishtik, et al. (2007). "Integrating process safety with molecular modeling-based risk assessment of chemicals within the REACH regulatory framework: benefits and future challenges." *J Hazard Mater* 142(3): 592-602.
- Lilienblum, W., W. Dekant, H. Foth, et al. (2008). "Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH)." *Archives of Toxicology* 82(4): 211-236.
- Linkov, I., P. Welle, D. Loney, et al. (2011). "Use of multicriteria decision analysis to support weight of evidence evaluation." *Risk Anal* 31(8): 1211-1225.
- Linkov, I., O. Massey, J. Keisler, et al. (2015). "From 'weight of evidence' to quantitative data integration using multicriteria decision analysis and Bayesian methods." *Altex* 32(1): 3-8.
- Luechtefeld, T., A. Maertens, J. M. McKim, et al. (2015). "Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships." *Journal of Applied Toxicology*: n/a-n/a.
- Luechtefeld, T., A. Maertens, D. P. Russo, et al. (2016). "Analysis of publically available skin sensitization data from REACH registrations 2008-2014." *Altex* 33(2): 135-148.
- MacKay, C., M. Davies, V. Summerfield, et al. (2013). "From pathways to people: applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment." *Altex* 30(4): 473-486.
- Margolin, B. H., K. J. Risko, M. D. Shelby, et al. (1984). "Sources of variability in Ames Salmonella typhimurium tester strains: Analysis of the International Collaborative Study on 'Genetic drift'." *Mutation Research/Environmental Mutagenesis and Related Subjects* 130(1): 11-25.
- Matheson, J., Zang, Q., Strickland, J., Kleinstreuer, N., Allen, D., Lowit, A., Jacobs, A., Casey, W. (2015). ICCVAM integrated decision strategy for skin sensitization. Extended abstract. 22-26 March in San Diego, USA.
- Mehling, A., T. Eriksson, T. Eltze, et al. (2012). "Non-animal test methods for predicting skin sensitization potentials." *Archives of Toxicology*: 1-23.
- Mekenyan, O., G. Patlewicz, G. Dimitrova, et al. (2010). "Use of Genotoxicity Information in the Development of Integrated Testing Strategies (ITS) for Skin Sensitization." *Chem Res Toxicol* 23(10): 1519-1540.
- Natsch, A., C. Bauch, L. Foertsch, et al. (2011). "The intra- and inter-laboratory reproducibility and predictivity of the KeratinoSens assay to predict skin sensitizers in vitro: Results of a ring-study in five laboratories." *Toxicology in Vitro* 25(3): 733-744.

- Natsch, A., C. A. Ryan, L. Foertsch, et al. (2013). "A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation." *Journal of Applied Toxicology* 33(11): 1337-1352.
- Natsch, A., R. Emter, H. Gfeller, et al. (2015). "Predicting skin sensitizer potency based on in vitro data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment." *Toxicol Sci* 143(2): 319-332.
- Nendza, M., M. Müller and A. Wenzel (2014). "Discriminating toxicant classes by mode of action: 4. Baseline and excess toxicity." *SAR and QSAR in Environmental Research* 25(5): 393-405.
- Nijsten, T. (2012). "Dermatology Life Quality Index: Time to Move Forward." *132*(1): 11-13.
- Nordberg, A., C. Rudén and S. O. Hansson (2008). "Towards more efficient testing strategies-Analyzing the efficiency of toxicity data requirements in relation to the criteria for classification and labelling." *Regulatory Toxicology and Pharmacology* 50(3): 412-419.
- Norlén, H., A. P. Worth and S. Gabbert (2014). "A tutorial for analysing the cost-effectiveness of alternative methods for assessing chemical toxicity: the case of acute oral toxicity prediction." *Alternatives to laboratory animals : ATLA* 42(2): 115-127.
- Nukada, Y., M. Miyazawa, S. Kazutoshi, et al. (2013). "Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals." *Toxicology in Vitro* 27(2): 609-618.
- OECD (1992). Test No. 406: OECD GUIDELINE FOR TESTING CHEMICALS. Adopted by the Council on 17th July 1992. Skin Sensitisation OECD Publishing.
- OECD (2002). "Skin Sensitisation: Local Lymph Node Assay. OECD Guideline for the Testing of Chemicals No 429, Paris. Available at: <http://www.oecd.org/env/testguidelines/>."
- OECD (2007). GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS Paris, OECD.
- OECD (2008). Workshop on Integrated Approaches to Testing and Assessment. OECD Series on Testing and Assessment No. 88, oECD.
- OECD (2010). Test No. 429: Skin Sensitisation: Local Lymph Node Assay, OECD Guidelines for the Testing of Chemicals, Section 4. Paris, OECD Publishing.
- OECD (2012a). OECD Environmental Outlook to 2050, OECD Publishing: 275-337.
- OECD (2012b). "The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Protein Part 1: Scientific Evidence." OECD Environment, health and safety publications No.168(ENV/JM/MONO(2012)10).
- OECD (2012c). "The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 2: Use of the AOP to Develop Chemical Categories and Integrated Assessment and Testing Approaches " OECD Environment, health and safety publications No 168(ENV/JM/MONO(2012)10/PART2).
- OECD (2012d). "<http://www.qsartoolbox.org/>."
- OECD (2015a). Test No. 442C: In Chemico Skin Sensitisation. Direct Peptide Reactivity Assay (DPRA), OECD Guidelines for the Testing of Chemicals, OECD Publishing, Paris.
- OECD (2015b). Test No. 442D: In Vitro Skin Sensitisation. In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD Guidelines for the Testing of Chemicals, OECD Publishing, Paris.
- OECD (2016a). PRODUCT RISK ASSESSMENT PRACTICES OF REGULATORY AGENCIES Summary of discussions at Workshops and Meetings of the OECD Working Party on Consumer Product Safety. Working Party on Consumer Product Safety. Paris, OECD publishing
- OECD (2016b). GUIDANCE DOCUMENT ON THE REPORTING OF DEFINED APPROACHES TO BE USED WITHIN INTEGRATED APPROACHES TO TESTING AND ASSESSMENT. Task Force on Hazard Assessment. France, OECD publishing
- OECD (2016c). GUIDANCE DOCUMENT ON THE REPORTING OF DEFINED APPROACHES AND INDIVIDUAL INFORMATION SOURCES TO BE USED WITHIN INTEGRATED APPROACHES TO TESTING AND ASSESSMENT (IATA) FOR SKIN SENSITISATION. France, OECD publishing
- OECD (2016d). Test No. 442E: In Vitro Skin Sensitisation: OECD GUIDELINE FOR THE TESTING OF CHEMICALS In Vitro Skin Sensitisation: human Cell Line Activation Test (h-CLAT). Paris, OECD Publishing.
- OECD (2016e). ANNEX I: CASE STUDIES TO THE GUIDANCE DOCUMENT ON THE REPORTING OF DEFINED APPROACHES AND INDIVIDUAL INFORMATION SOURCES TO BE USED WITHIN INTEGRATED APPROACHES TO TESTING AND ASSESSMENT (IATA FOR SKIN SENSITISATION). S. o. T. A. N. 256. Paris, OECD publishing
- Olson, L. J. (1990). "The search for a safe environment: The economics of screening and regulating environmental hazards." *Journal of Environmental Economics and Management* 19(1): 1-18.
- Omenn, G. S. (1995). "Assessing the risk assessment paradigm." *Toxicology* 102(1-2): 23-28.
- Ostaszewski K. and Rempala G.A. (2000). "Parametric and Nonparametric Bootstrap in Actuarial Practice." University of Louisville.
- Paparella, M., M. Daneshian, R. Hornek-Gausterer, et al. (2013). "Uncertainty of testing methods--what do we (want to) know?" *Altex* 30(2): 131-144.
- Park, M. E. and J. H. Zippin (2014). "Allergic contact dermatitis to cosmetics." *Dermatologic Clinics* 32(1): 1-11.
- Patlewicz, G., C. Kuseva, A. Kesova, et al. (2014). "Towards AOP application – Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization." *Regulatory Toxicology and Pharmacology* 69(3): 529-545.
- Patlewicz, G., T. W. Simon, J. C. Rowlands, et al. (2015). "Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes." *Regulatory Toxicology and Pharmacology* 71(3): 463-477.
- Peiser, M., T. Tralau, J. Heidler, et al. (2012). "Allergic contact dermatitis: epidemiology, molecular mechanisms, in vitro methods and regulatory aspects: Current knowledge assembled at an international workshop at BfR, Germany." *Cellular and Molecular Life Sciences* 69(5): 763-781.
- Perkins, E. J., P. Antczak, L. Burgoon, et al. (2015). "Adverse outcome pathways for regulatory applications: Examination of four case studies with different degrees of completeness and scientific confidence." *Toxicological Sciences* 148(1): 14-25.

References

- Ramirez, T., A. Mehling, S. N. Kolle, et al. (2014). "LuSens: A keratinocyte based ARE reporter gene assay for use in integrated testing strategies for skin sensitization hazard identification." *Toxicology in Vitro* 28(8): 1482-1497.
- Ramirez, T., N. Stein, A. Aumann, et al. (2016). "Intra- and inter-laboratory reproducibility and accuracy of the LuSens assay: A reporter gene-cell line to detect keratinocyte activation by skin sensitizers." *Toxicology in Vitro* 32: 278-286.
- Reisinger, K., S. Hoffmann, N. Alépée, et al. (2015). "Systematic evaluation of non-animal test methods for skin sensitisation safety assessment." *Toxicology in Vitro* 29(1): 259-272.
- Ricci, G., B. Bendandi, L. Pagliara, et al. (2006). "Atopic Dermatitis in Italian Children: Evaluation of Its Economic Impact." *Journal of Pediatric Health Care* 20(5): 311-315.
- Rorije, E., T. Aldenberg, H. Buist, et al. (2013). "The OSIRIS Weight of Evidence approach: ITS for skin sensitisation." *Regulatory Toxicology and Pharmacology* 67(2): 146-156.
- Rovida, C. and T. Hartung (2009). "Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t(4))." *Altex* 26(3): 187-208.
- Rovida, C., N. Alépée, A. M. Api, et al. (2015). "Integrated testing strategies (ITS) for safety assessment." *Altex* 32(1): 25-40.
- Russell, W. M. S. a. B., R. L. (1959). "The Principles of Humane Experimental Technique." 1st ed., London, UK: Methuen & Co Ltd. ed. Universities Federation for Animal Welfare.
- Sætterstrøm, B., J. Olsen and J. D. Johansen (2014). "Cost-of-illness of patients with contact dermatitis in Denmark." *Contact Dermatitis* 71(3): 154-161.
- Sakaguchi, H., T. Ashikaga, M. Miyazawa, et al. (2006). "Development of an in vitro skin sensitization test using human cell lines; human Cell Line Activation Test (h-CLAT) II. An inter-laboratory study of the h-CLAT." *Toxicology in Vitro* 20(5): 774-784.
- Sakaguchi, H., C. Ryan, J. M. Ovigne, et al. (2010). "Predicting skin sensitization potential and inter-laboratory reproducibility of a human Cell Line Activation Test (h-CLAT) in the European Cosmetics Association (COLIPA) ring trials." *Toxicology in Vitro* 24(6): 1810-1820.
- Sauer, U. G., E. H. Hill, R. D. Curren, et al. (2016). "Local tolerance testing under REACH: Accepted non-animal methods are not on equal footing with animal tests." *Alternatives to laboratory animals : ATLA* 44(3): 281-299.
- Schaafsma, G., E. D. Kroese, E. L. J. P. Tielemans, et al. (2009). "REACH, non-testing approaches and the urgent need for a change in mind set." *Regulatory Toxicology and Pharmacology* 53(1): 70-80.
- Schnuch, A., J. Geier, W. Uter, et al. (1997). "National rates and regional differences in sensitization to allergens of the standard series." *Contact Dermatitis* 37(5): 200-209.
- Schnuch, A., H. Lessmann, J. Geier, et al. (2011). "Contact allergy to preservatives. Analysis of IVDK data 1996-2009." *British Journal of Dermatology* 164(6): 1316-1325.
- Schnuch, A., J. Geier, H. Lessmann, et al. (2012). "Surveillance of contact allergies: methods and results of the Information Network of Departments of Dermatology (IVDK)." *Allergy* 67(7): 847-857.
- Schoeters, G. (2010). "The reach perspective: Toward a new concept of toxicity testing." *Journal of Toxicology and Environmental Health - Part B: Critical Reviews* 13(2-4): 232-241.
- Schultz, T. W., G. Dimitrova, S. Dimitrov, et al. (2016). "The adverse outcome pathway for skin sensitisation: Moving closer to replacing animal testing." *Alternatives to laboratory animals : ATLA* 44(5): 453-460.
- Simon, T. W., S. S. Simons, R. J. Preston, et al. (2014). "The use of mode of action information in risk assessment: Quantitative key events/dose-response framework for modeling the dose-response for key events." *Crit Rev Toxicol* 44(S3): 17-43.
- Slob, W. (2016). "A general theory of effect size, and its consequences for defining the benchmark response (BMR) for continuous endpoints." *Crit Rev Toxicol*: 1-10.
- Stein, V., T. Dorner, K. Lawrence, et al. (2007). "Economic aspects of allergies: Status and prospects for Austria." *Wiener Medizinische Wochenschrift* 157(11-12): 248-254.
- Strickland, J., Q. Zang, N. Kleinstreuer, et al. (2016). "Integrated decision strategies for skin sensitization hazard." *Journal of Applied Toxicology*: n/a-n/a.
- Sullivan, K. (2016). "It takes a village: Stakeholder participation is essential to transforming science." *ATLA Alternatives to Laboratory Animals* 44(5): 411-415.
- Takenouchi, O., S. Fukui, K. Okamoto, et al. (2015). "Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals." *J Appl Toxicol* 35(11): 1318-1332.
- Thyssen, J. P., A. Linneberg, T. Menne, et al. (2007). "The epidemiology of contact allergy in the general population - prevalence and main findings." *Contact Dermatitis* 57(5): 287-299.
- Tollefsen, K. E., S. Scholz, M. T. Cronin, et al. (2014). "Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA)." *Regulatory Toxicology and Pharmacology* 70(3): 629-640.
- Tralau, T., M. Oelgeschläger, R. Gürtler, et al. (2015). "Regulatory toxicology in the twenty-first century: challenges, perspectives and possible solutions." *Archives of Toxicology* 89(6): 823-850.
- Tsujita-Inoue, K., M. Hirota, T. Ashikaga, et al. (2014). "Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays." *Toxicology in Vitro* 28(4): 626-639.
- Tsujita-Inoue, K., T. Atobe, M. Hirota, et al. (2015). "In silico risk assessment for skin sensitization using artificial neural network analysis." *J Toxicol Sci* 40(2): 193-209.
- UNECE (2011). Chapter 3.4 Respiratory or skin sensitization. GLOBALLY HARMONIZED SYSTEM OF CLASSIFICATION AND LABELLING OF CHEMICALS (GHS), Fourth revised version
New York and Geneva United Nations
- Urbisch, D., A. Mehling, K. Guth, et al. (2015a). "Assessing skin sensitization hazard in mice and men using non-animal test methods." *Regulatory Toxicology and Pharmacology* 71(2): 337-351.
- Urbisch, D., Honarvar N., Mehling A., et al. (2015b). Regulatory Use of Non-Animal Test Methods in Chemical Industry: The Example of Skin Sensitization. Towards the Replacement of in vivo Repeated Dose Systematic Toxicity Testing. t. E. C.

- SEURAT-1 Initiative, and the European Cosmetics Association (Cosmetics Europe). Printed in France - Imprimerie Mouzet: 42-54.
- Urbisch, D., M. Becker, N. Honarvar, et al. (2016). "Assessment of Pre- and Pro-haptens Using Nonanimal Test Methods for Skin Sensitization." *Chem Res Toxicol* 29(5): 901-913.
- Uter, W., O. Gefeller, J. Geier, et al. (2012). "Methylchloroisothiazolinone/methylisothiazolinone contact sensitization: Diverging trends in subgroups of IVDK patients in a period of 19 years." *Contact Dermatitis* 67(3): 125-129.
- Uter, W., J. Geier, A. Bauer, et al. (2013). "Risk factors associated with methylisothiazolinone contact sensitization." *Contact Dermatitis* 69(4): 231-238.
- Valérie Zuang, B. D., João Barroso, Susanne Belz, Elisabet Berggren, Camilla Bernasconi, Jos Bessems, Stephanie Bopp, Silvia Casati, Sandra Coecke, Raffaella Corvi, Coralie Dumont, Varvara Gouliarmou, Claudius Griesinger, Marlies Halder, Annett Janusch-Roi, Aude Kienzler, Brigitte Landesmann, Federica Madia, Anne Milcamps, Sharon Munn, Anna Price, Pilar Prieto, Michael Schäffer, Jutta Triebe, Clemens Wittwehr, Andrew Worth and Maurice Whelan (2015). EURL ECVAM Status Report on the Development, Validation and Regulatory Acceptance of Alternative Methods and Approaches (2015). Luxembourg, Publications Office of the European Union, 2015.
- van der Schouw, Y. T., Verbeek A. L., and S. H. Ruijs (1995). "Guidelines for the assessment of new diagnostic tests." *Invest Radiology* 30(6): 334-340.
- van der Veen, J. W., E. Rorije, R. Emter, et al. (2014a). "Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals." *Regulatory Toxicology and Pharmacology* 69(3): 371-379.
- van der Veen, J. W., L. G. Soeteman-Hernandez, J. Ezendam, et al. (2014b). "Anchoring molecular mechanisms to the adverse outcome pathway for skin sensitization: Analysis of existing data." *Crit Rev Toxicol* 44(7): 590-599.
- van Leeuwen, C. J., T. G. Vermeire, C. J. Leeuwen, et al. (2007). *Intelligent Testing Strategies. Risk Assessment of Chemicals*, Springer Netherlands: 467-509.
- Verboom, P., L. Hakkaart-van Roijen, M. Sturkenboom, et al. (2002). "The cost of atopic dermatitis in the Netherlands: an international comparison." *British Journal of Dermatology* 147(4): 716-724.
- Vermeire, T., T. Aldenberg, H. Buist, et al. (2013). "OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints." *Regul Toxicol Pharmacol* 67(2): 136-145.
- Villeneuve, D., D. C. Volz, M. R. Embry, et al. (2014). "Investigating Alternatives to the fish early-life stage test: A strategy for discovering and annotating adverse outcome pathways for early fish development." *Toxicology* 33(1): 158-169.
- Vinken, M. (2013). "The adverse outcome pathway concept: A pragmatic tool in toxicology." *Toxicology* 312(0): 158-165.
- Vinken, M., M. Whelan and V. Rogiers (2014). "Adverse outcome pathways: hype or hope?" *Archives of Toxicology* 88(1): 1-2.
- Vonk, J. A., R. Benigni, M. Hewitt, et al. (2009). "The use of mechanisms and modes of toxic action in integrated testing strategies: the report and recommendations of a workshop held as part of the European Union OSIRIS Integrated Project." *Altern Lab Anim* 37(5): 557-571.
- Wallace, A. D. and H. Ernest (2012). Chapter Four - Toxic Endpoints in the Study of Human Exposure to Environmental Chemicals. *Progress in Molecular Biology and Translational Science*, Academic Press. Volume 112: 89-115.
- Weil, C. S. and R. A. Scala (1971). "Study of intra- and interlaboratory variability in the results of rabbit eye and skin irritation tests." *Toxicology and Applied Pharmacology* 19(2): 276-360.
- Wehrens, R., H. Putter and L. M. C. Buydens (2000). "The bootstrap: a tutorial." *Chemometrics and Intelligent Laboratory Systems* 54(1): 35-52.
- WHO (2004). Harmonization Project Document No. 1 IPCS RISK ASSESSMENT TERMINOLOGY This project was conducted within the IPCS project on the Harmonization of Approaches to the Assessment of Risk from Exposure to Chemicals. Geneva.
- WHO (2016). The public health impact of chemicals: knowns and unknowns. Geneva.
- WHO/FAO (2016a). International code of conduct on pesticide management: guidelines on highly hazardous pesticides. Rome Food and Agriculture Organization of the United Nations.
- WHO/FAO (2016b). Manual on development and use of FAO and WHO specifications for pesticides. P. SPECIFICATIONS. Geneva and Rome, Joint Meeting on Pesticide Specifications (JMPS).
- Witt, C. M., B. Brinkhaus, D. Pach, et al. (2009). "Homoeopathic versus Conventional Therapy for Atopic Eczema in Children: Medical and Economic Results." *Dermatology* 219(4): 329-340.
- Worth, A., J. Barroso, S. Bremer, et al. (2014). "Alternative methods for regulatory toxicology - a state-of-the-art review." Publications Office of the European Union.
- Worth, A. P. and M. T. D. Cronin (2001a). "The use of bootstrap resampling to assess the uncertainty of Cooper statistics." *Atla-Alternatives to Laboratory Animals* 29(4): 447-459.
- Worth, A. P. and M. T. D. Cronin (2001b). "The use of bootstrap resampling to assess the variability of Draize tissue scores." *ATLA Alternatives to Laboratory Animals* 29(5): 557-573.
- Worth, A. P. and G. Patlewicz (2016). Integrated approaches to testing and assessment. *Advances in Experimental Medicine and Biology*. 856: 317-342.
- Wynand, P. M., M. V. De Ven, R. P. Ellis, et al. (2000). Chapter 14 - Risk Adjustment in Competitive Health Plan Markets*. *Handbook of Health Economics*, Elsevier. Volume 1, Part A: 755-845.
- Yerushalmy, J. (1947). "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques." *Public Health Rep* 62(40): 1432-1449.
- Yokota, F., G. Gray, J. K. Hammit, et al. (2004). "Tiered chemical testing: A value of information approach." *Risk Analysis* 24(6): 1625-1639.
- Yokota, F. and K. M. Thompson (2004). "Value of Information Analysis in Environmental Health Risk Management Decisions: Past, Present, and Future." *Risk Analysis* 24(3): 635-650.
- Zaubar "<https://www.zaubar.com/importanalysis-kathon-report.html>", (Assessed 21.07.2016)

The research described in this thesis was financially supported by BASF SE, Germany.

Cover design: Prepress Manager & Graphic Designer, Wageningen, NL.

Printed by: Prepress Manager & Graphic Designer, Wageningen, NL.



*Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment*

D I P L O M A

For specialised PhD training

The Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Maria Leontaridou

born on 12 April 1983 in Athens, Greece

has successfully fulfilled all requirements of the
Educational Programme of SENSE.

Wageningen, 19 April 2017

the Chairman of the SENSE board

Prof. dr. Huub Rijnaarts

the SENSE Director of Education

Dr. Ad van Dommelen

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)



K O N I N K L I J K E N E D E R L A N D S E
A K A D E M I E V A N W E T E N S C H A P P E N



The SENSE Research School declares that **Ms Maria Leontaridou** has successfully fulfilled all requirements of the Educational PhD Programme of SENSE with a work load of 60.1 EC, including the following activities:

SENSE PhD Courses

- o Environmental research in context (2013)
- o Research in context activity: "Co-organising SENSE summer symposium : Make a change! successful interaction with society in sustainability science" (2015)

Other PhD and Advanced MSc Courses

- o Scientific writing, Wageningen University (2012)
- o Presentation skills, Wageningen University (2012)
- o Decision science 1, Wageningen University (2012)
- o General toxicology, Wageningen University (2013)
- o Environmental economics for environmental sciences, Wageningen University (2013)
- o Quantitative research methodology and statistics, Wageningen University (2013)
- o Cost-benefit analysis and environmental valuation, Wageningen University (2013)
- o Advanced microeconomics, Wageningen University (2014)

External training at a foreign research institute

- o Six months' scientific visit, Institute for Health and Consumer Protection (IHCP) - Joint Research Institute (JRC), Ispra, Italy (2014)
- o Three months' scientific visit, BASF Department of Experimental Toxicology and Ecology, Ludwigshafen, Germany (2015)

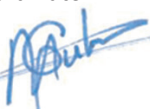
Management and Didactic Skills Training

- o Member of the Wageningen PhD student Council (WPC) (2013-2014)
- o Member of the WIMEK PhD student Council (2014-2015)

Oral Presentations

- o *Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian value-of-information analysis.* European Society for Alternatives to Animal Testing (EUSAT), 20-23 September 2015, Linz, Austria
- o *The economics of toxicity testing.* Conference Developing Integrated Multi-level Models of the Environment, 7-9 March 2016, Liverpool, United Kingdom

SENSE Coordinator PhD Education



Dr. ing. Monique Gulickx