(Q)SAR and other (non) testing data in integrated testing strategies using standardised weight of evidence criteria
Hulzebos, E.M.

Citation for published version (APA):
Hulzebos, E. M. (2012). (Q)SAR and other (non) testing data in integrated testing strategies using standardised weight of evidence criteria

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 1

General Introduction

Etje Hulzebos
Introduction

Modern society produces more and more chemicals in increasing volumes. The impact of these chemicals on health and environment needs to be investigated to guarantee their safe use. Traditionally, animal testing has been the main approach to assess this safety. However, concern in the society on the use of animals in bio- and medical research and testing is increasing. This resulted in a framework of political and legislative measures. Many successful attempts have been made to decrease the number of animals used and to increase animal welfare.

The first legislation on the use of animals in research and testing in the Netherlands was the Animals Experimentation Act (Wet op de Dierproeven) in 1977. This law required registration and approval of animal experiments. Thereafter, Council Directive 86/609/EC came into force in 1986 in the Member States of the EU. This Directive was revised in the last decade and adopted by European Parliament in 2010 (EC-Directive 2010/63/EC, 2010) and will be implemented in Member State laws in 2013. Another successful contribution to limit animal use and to make testing more humane is the establishment of international standardisation of (animal) testing methods for toxicity, such as the Testing Guidelines of the Organisation for Economic Co-operation and Development (OECD TG, 2011). The information generated according to these methods is accepted throughout the world and thus no repeat testing is needed to accommodate country specific requirements.

The initiatives to endorse humane strategies into animal testing can be clustered into the so called 3Rs Principle (Russel and Burch, 1959 and e.g. Hartung, 2006):

1) Replace: to avoid the use of vertebrate animals altogether by using non-animal methods or non-vertebrates. Both the European Centre for the Validation of Alternative Methods (ECVAM) and the American Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) play a prominent role in coordinating the validation of in vitro and in vivo testing methods (ECVAM and ICCVAM, 2011). Next to testing methods also non-testing methods, so called in silico methods, have been propagated for example analogue approaches and (quantitative) structure-activity relationships ((Q)SARs), which will be discussed later. The coordination and promotion of such in silico testing has been done under the umbrella of the OECD and at the Joint Research Centre (JRC) in Ispra (OECD-QSAR Project, 2011, JRC-Computational Toxicology, 2011).

The focus of this thesis is Replacement either by using non-animal tests or approaches as a standalone method, or as part of an Integrated Testing Strategy (ITS), ultimately resulting in the use of fewer animals.

2) Reduce: obtain information from as few as possible number of animals for answering the scientific question or obtaining more information.
from the same number of animals. The reduction of animals as described here is outside the scope of this thesis.

3) Refine: make sure animals suffer as little as possible or increase the welfare of the animals. Refinement of testing is also not part of this thesis.

Currently, there is a wide range of research initiatives with a focus on developing new tests and test strategies to limit animal use. These have been funded by (inter)national public and private organisations. Examples are given below:

- Ministry of Infrastructure and Environment and Ministry of Public Health funded via the organisation ZonMw projects on acute oral toxicity, skin irritation, skin sensitisation, and reproductive toxicity (Freidig et al., 2007, Hulzebos et al., 2005 a and b, Fabjan and Hulzebos, 2008, Maslankiewicz et al. 2005, respectively). More recently ZonMw (2011) funded a project on implementing alternatives for repeated dose toxicity (e.g. Punt et al., 2011);

- European Commission: There are two running EU framework programmes funding research into alternatives to animal testing, which includes (Q)SARs, which will be mentioned here. The first one is Cadaster, which focuses on (Q)SARs for aquatic toxicity and Persistent Bioaccumulation and Toxicity (PBT) profiling (Cadaster, 2011). The second one is the OSIRIS (2011) project, which focuses on ITSs. One of its products is the publication of Vonk et al. (2009), who focus on mechanisms and modes of toxic action in ITSs.

- Industry in their International Council of Chemical Associations: The Long-Range Research Initiative (ICCA/LRI) initiated (Q)SAR research, which resulted in many scientific publications on this subject. Ecotoxicological (Q)SARs were reviewed in 1998 and human health (Q)SARs in 2003 (ECETOC, 1998, 2003). In 2003 an OECD workshop was co-organised to define criteria for (Q)SAR application for regulatory purposes (OECD, 2004). In 2010 ECETOC evaluated new non-animal testing methods using high-throughput screening to support approaches using analogue chemicals (ECETOC, 2010).

In Europe the current approach for risk assessment is using animal testing for (eco)toxicological effect assessment. A new policy of risk assessment of industrial chemicals in Europe is laid down in the REACH-Regulation, which stands for Registration, Evaluation, Authorisation and restriction of Chemical substances (EC, 2006). In this regulation the replacement of animal testing is feasible under strict criteria. In Annex XI of this regulation it is laid out that information required may be derived from testing and non-testing approaches (EC, 2006, and in the amendment of 2010). The Cosmetic Directive has taken a more rigorous approach and wants to replace all animal testing by 2013 for cosmetic products and their ingredients (EC-Cosmetic Directive, 2009).
In Europe and in the US a new vision and strategy is being developed for the risk assessment of chemicals (ASAT, 2008 and NRC, 2007). This new risk assessment focuses on testing methods identifying effects on the molecular and cellular level, which subsequently are extrapolated to a potential adverse effect in humans.

The aim of this thesis

Though the possibility of using alternatives is presented in the REACH legislation, the application of these methodologies can cause cold feet to regulatory toxicologists in absence of internationally harmonised and validated methods and limited guidance. To bridge the gap between scientific work on these alternative methods and their regulatory application much research has been carried out during the last ten years. This thesis is a reflection of this research. More specifically, the purpose of this thesis is to investigate whether (Q)SARs can be applied successfully within a regulatory framework such as REACH and, if so, under what conditions animal testing replacement is feasible.

In line with this European policy the purpose of this thesis is to contribute to the replacement of animal testing while ensuring the safe use of chemicals. The starting point is to investigate the application of in silico methods for regulatory purposes, such as (Q)SARs, which can predict (eco)toxicity based on chemical structure. Chapters 2-4 focus on:

1) The predictivity of (Q)SARs methods for (eco)toxicological endpoints (Chapter 2, Hulzebos and Posthumus, 2003);
2) The evaluation of these (Q)SAR tools according to the OECD principles for the validation of (Q)SARs (Chapter 3, Hulzebos et al., 2005a);
3) Developing a hybrid SAR skin irritation tool (Chapter 4, Hulzebos et al., 2005b).

In this thesis it is argued that (Q)SARs can be used, often not as a standalone method, but as part of an ITS for assessing ecotoxicological and human health toxicity to be referred to as (eco)toxicity. In an ITS all available data are taken into account and weighted to determine the hazard of a chemical. Such data may be existing animal data (in vivo) or non-animal data such as (Q)SARs and in vitro data. By themselves, each of them may give insufficient information for a final decision on the (eco)toxicity of a chemical. However when these data are weighted and combined a more complete picture on the hazard of the tested chemical can come forward and a more reliable decision can be made. Therefore a novel tool has been developed:

4) the Integrated Assessment Scheme (IAS). This tool is expected to guide the risk assessor in the evaluation and weighting of information from (Q)SARs and other (eco)toxicological information, using a similar denominator. It aims to help the risk assessor in the decision making and the choice of which information to use for a particular regulatory
purpose. It can also help in the assessment whether information can be applied as standalone, as an information element in a testing strategy, or should not be used (Chapter 5, Hulzebos et al., 2010);

5) the application of this IAS will be shown for the assessment of skin irritation using (Q)SAR predictions and in vivo and in vitro data (Chapter 6, Hulzebos and Gerner, 2010).

This scheme is of utmost importance to weigh and evaluate the (eco)toxicity data of chemicals in the integration of data from different methods and for different purposes (Hulzebos et al., 2010 and Hulzebos and Gerner, 2010). It will be shown that the expert judgment, indispensible needed for weighting and evaluation of data, can now be documented in a more objective, formalised, transparent and consistent way (ECHA, 2010).

First the process of risk assessment of chemicals for the regulatory application will be introduced. This will be followed by a literature review introducing (Q)SARs and ITSs. Finally some remarks will be made on the structure of this thesis.

Risk assessment of chemicals

General introduction

Chemical risk assessment is meant to show that the use of a chemical does or does not cause concern for humans and the environment. A risk assessment on chemicals can be captured in a three step process: hazard assessment (including hazard identification), exposure assessment and risk characterisation (Van Leeuwen, 2007, Fig. 1).

Hazard assessment is the first step (Step 1). Hazard is the adverse effect of a chemical as an inherent property. It is the likelihood of harm due to exposure that distinguishes risk from hazard. For example, a toxic chemical that is hazardous to human health does not constitute a risk unless humans are exposed to it.

The hazards in humans may include acute toxicity, skin and eye irritation, sensitisation, systemic effects, reproductive toxic effects or cancer (human health endpoints). Ecological hazards include lethal effects such as mortality and sub lethal effects on growth and reproduction of various populations (ecotoxicological endpoints). For the ecological hazards also fate parameters are included, such as ‘ready biodegradability’ to assess the persistency in the environment and the bioaccumulation potential to assess the accumulation of the chemical in the food chain.
A first identification on the chemical’s hazard can often be found in its chemical structure and its physico-chemical properties. In the hazard assessment the adversity of the effects has to be defined. Secondly the dose response assessment observed in animals tested in the laboratory and all other available data e.g. human data will be included in this assessment.

The dose response assessment is the estimation of the relationship between dose or level of direct exposure to a chemical and the incidence and severity of an effect. The data may be obtained from (Q)SARs, analogue approaches, in vitro studies, experimental plant and laboratory animal studies (in vivo studies) or a combination of these data evaluated in an ITS.

From these effect data a No Observed Adverse Effect Level (NOAEL) or a No Observed Effect Concentration (NOEC) is derived for the species that is tested. A NOAEL in repeated dose testing is the highest experimental dose without adverse effect. A decision on the adversity of the effect includes the severity and frequency of the effects seen and the reversibility of these effects when the exposure to the chemical ends. Whether an effect is considered adverse is sometimes debatable. Doubtful relevance of some toxicological effects in a repeated dose study of 28-days may be considered adverse unless it is proven not relevant in a repeated dose study of 90-days.

Similarly, for ecotoxicity for acute effects LC50 and EC50 values are derived. The LC50 is the (calculated) concentration causing 50% mortality in fish. The EC50 is the (calculated) concentration causing 50% immobilisation in Daphnia or 50% growth inhibition in an algae test. A NOEC is the highest concentration without observed effects.

To derive a safe level or concentration for humans and the environment, assessment factors are used to account for uncertainty due to differences between and within species and between short- and long-term exposure. The interspecies differences include different metabolic and biological pathways between humans and animals. Within REACH, safe levels are called Derived No Effect levels (DNELs) for human health and Predicted No Effect Concentrations (PNECs) for the environment. After deriving these DNELs and PNECs, information on potential exposure to humans and environment is needed to see whether risks are encountered (Step 2).
Fig. 1 Illustration of the risk assessment process (revised from van Leeuwen, 2007).

Step two is the exposure assessment. This involves estimating releases and the distribution of the chemical to define the doses (levels) or concentrations to which human and environmental populations are exposed. The human populations considered are workers (during manufacturing, formulation of the chemical or use of end products containing it), consumers (using consumer products) and humans exposed via the environment (e.g. exposure via drinking water). Exposure in the environment occurs through the compartments air, water, sediment and soil. Due to high variability in human and environmental exposures and the absence of representative monitoring data, exposure assessment is often based on modelling. These calculated exposures are generally referred to as Predicted Exposure Levels (PELs) or Predicted Exposure Concentrations (PECs). These exposure levels and concentrations have the same units as those of the effect assessment and thus can be compared to evaluate potential risks, which is discussed in the third step, the risk characterisation.
The risk characterisation is the actual estimation of the incidence and severity of the hazards likely to occur in human populations or environmental compartments, due to actual or predicted exposure to a chemical. Under REACH the PELs and PECs are compared with the DNELs and the PNECs to estimate these adverse effects based on a ‘realistic worst case scenario’. When the PEL/DNEL < 1 or the PEC/PNEC is < 1 the risk is considered to be not significant. When a risk quotient is higher than one, further information is needed to refine the risk characterisation. This can be done by gathering more information on exposure or initiate further (long-term) testing which may be a reason to lower the assessment factors which are being used to extrapolate from e.g. short to long-term testing.

The risk assessments of chemicals for regulatory purposes are categorised according to their main application in society, e.g. industrial chemicals, pharmaceuticals, pesticides, veterinary drugs and food additives. Table 1 presents an overview of the main regulatory frameworks for assessing chemicals and their scientific committees at the level of European Commission. It is extracted from van Leeuwen enVermeire (2007, Chapter 8).

There is considerable overlap between the different regulatory frameworks requiring testing information to assess the hazards of chemicals. For hazard information they all require that new testing is carried out using OECD TGs in compliance with Good Laboratory Practice (GLP). This is to make sure that the information is acceptable for regulatory authorities all over the world. The focus of the pharmaceutical assessment for human safety will mostly be on the individual’s wellbeing, making sure that the medicinal effect clearly outweighs the side effects. For pesticides, besides being effective, the exposure of the worker who applies the pesticide is looked into and for industrial chemicals the worker’s exposure during the whole life cycle is assessed. Consumer use and exposure are also assessed. Most regulatory frameworks now also require an ecotoxicological risk assessment.

**Risk assessment of chemicals before REACH**

New Chemical legislations came into force after a number of disasters caused by chemicals. Examples are the Love Canal toxic waste site, the Seveso disaster in 1976, causing the highest known exposure of TCDD in urban populations and the Bhopal disaster in 1984 in which leakage of methylisocyanate in the pesticides plant of Union Carbide caused poisoning and death of countless people. It became evident that many substances and their metabolites were persistent in the environment and bioaccumulated, causing major problems for humans, animals and the environment. In addition, for workers exposure to certain chemicals may have a great impact. Dioxins for example became known carcinogens. Because of the impact of these industrial chemicals on human health and the environment, regulatory authorities all over the world have put
legislation in place to prevent similar situations as described above. In these legislations data need to be submitted for newly marketed chemicals. In this way it can be shown that the new chemical is safe for humans and the environment.

Table 1  Types of chemicals and scientific committees involved in risk assessment (van Leeuwen en Vermeire, 2007, Chapter 8) and *updated in 2012

<table>
<thead>
<tr>
<th>Types of chemical</th>
<th>Directive number</th>
<th>Organisation</th>
<th>Scientific Committee (SC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industrial chemicals</td>
<td>EC 1907/2006 and 2006/121/EC</td>
<td>European Chemicals Agency (ECHA)</td>
<td>Not established yet</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>2004/27/EC</td>
<td>European Medicines Agency (EMA)</td>
<td>SC for Medicinal Products and human use: CHMP</td>
</tr>
<tr>
<td>Veterinary medicines</td>
<td>2001/82/EC</td>
<td>European Medicines Agency (EMA)</td>
<td>SC for Medicinal Products and Veterinary use: CCMP</td>
</tr>
<tr>
<td>Residues of veterinary medicines and additives*</td>
<td>470/2009/EC and 1831/2003/EC</td>
<td>European Food Safety Authority (EFSA)</td>
<td>Panel on Additives and Products or Substances used in Animal Feed (FEEDAP)</td>
</tr>
<tr>
<td>Pesticides</td>
<td>91/414/EC</td>
<td>European Food Safety Authority (EFSA)</td>
<td>Panel on Plant Protection and their residues: PPR</td>
</tr>
<tr>
<td>Food additives</td>
<td>89/107/EEC</td>
<td>European Food Safety Authority (EFSA)</td>
<td>Panel on Food Additives and Nutrient Sources Added to Food (ANS)</td>
</tr>
<tr>
<td>Food contact materials</td>
<td>2007/42/EC</td>
<td></td>
<td>Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF)</td>
</tr>
</tbody>
</table>


The update of the New Chemical legislation in Europe was followed up by the Existing Chemical regulation in 1992 in Europe (EC, 1992 and 1993, respectively) partly to meet the criticism that for high volume chemicals limited data were available while for a low tonnage new chemicals more data needed to be submitted. Priority tools were developed to select the high volume chemicals with a potential hazard and 140 chemicals were selected (Bodar et al., 2002). The Existing Chemical assessment proved to be very time consuming and complex: of the 140 chemicals selected 70 risk assessment reports were finalised in 2007 which has increased to 137 up to 2012 (Wormuth et al., 2007, JRC-ORATS, 2012). This slow process was caused by the additional
experimental testing, that often needed to be done, but also by the time consuming harmonisation and consensus building with regard to the risk assessment between the EU member states (Bodar et al., 2002). After these eight years about 30,000 chemicals, which are marketed above one tonnage, still needed to be evaluated. The need for the evaluation of these chemicals among other reasons required a new policy, which was laid down in the White Paper (EC, 2001). In this White Paper one of the objectives of the new policy was to document the safety assessment of chemicals. In order to save time and money the responsibility of doing risk assessment was given to industry. In addition, there was a strong emphasis on using alternative test strategies and methods, such as (Q)SARs, analogue approaches, in vitro testing and testing strategies, to limit animal testing (EC, 2001). The White Paper reflected the growing concern about the extensive use of animals for testing, which became more and more criticized ever since the 1960’s. This new policy of risk assessment of industrial chemicals in Europe is now laid down in the REACH regulation (EC, 2006).

**Risk assessment of chemicals under REACH**

The REACH regulation not only integrates the two former legislations on New and Existing (industrial) chemical regulations but also replaces 40 existing legal acts and creates a single system for all chemical substances (EC, 2006). A tool to prevent animal testing under REACH is the Substance Information Exchange Forums (SIEFs). Amongst other reasons, these Forums have been established to bring registrants of chemicals into contact with each other to share animal data they may have. To promote the use of alternative testing the non-animal testing methods have become part of the risk assessment process in the EU under the REACH regulation. In Annex XI of this legislation it is explained how data may be derived from testing and non-testing approaches to fulfil the information requirements (EC, 2006, and in the amendment of 2010). Another way to limit animal testing under REACH is the requirement for submitting testing proposals for studies on vertebrate animals for marketed tonnages > 100 tons such as the longer term toxicity tests e.g. 90-day study, developmental toxicity study and bioaccumulation study. These testing proposals are put on the European Chemicals Agency (ECHA) website inviting others to forward data they may have. In addition, one can comment on the need to perform such animal studies (ECHA, 2011a).

Under REACH, risk assessment consists of the risk characterisation described above. It also requires Classification, Labelling and Packaging according to the CLP regulation (EC-CLP, 2008). According to Annex XIII of the regulation the Persistent, Bioaccumulation and Toxicity (PBT) assessment has to be performed. These assessments focus on the hazard and fate assessment (e.g. biodegradability and bioaccumulation) of chemicals. The REACH information requirements depend on the marketed tonnages. In Annex VII to Annex X the
data requirements are laid out on identity, hazard and exposure, which needs to be submitted by industry when the marketed tonnage is > 1 tons. The information has to be entered into the IUCLID (International Uniform Chemical Information Database) and has to be submitted to the ECHA.

An overview of the type of data that needs to be submitted for REACH when the marketing is above 10 tons per annum is presented in Table 2 and involves extensive animal testing. A full exposure assessment is needed for chemicals marketed for > 10 tons and which are classified and labelled. More long-term testing is required for marketed tonnage > 100 tons but this often depends on the results of the information gathered at the > 10 tons level.

<table>
<thead>
<tr>
<th>REACH requirements</th>
<th>Human health endpoints</th>
<th>Environmental endpoints</th>
<th>REACH purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazard information (Step 1)</td>
<td>- Acute toxicity - Skin and eye irritation - Sensitisation - Repeated dose toxicity - Reproductive toxicity - Genotoxicity</td>
<td>- Algae, Daphnia and Fish toxicity - Ready biodegradability - Bioaccumulation potential</td>
<td>- DNELs and PNECs - Classification and Labelling - PBT assessment</td>
</tr>
<tr>
<td>Exposure information (Step 2)</td>
<td>- Worker, - Consumers - Man exposed via the environment</td>
<td>- Fresh water - Sea water - Sediment organisms - Terrestrial organisms</td>
<td>Risk characterisation - DNEL vs. Exposure - PNEC vs. Exposure</td>
</tr>
</tbody>
</table>

**Assigning quality to data**

Assessing the data quality is a key process when submitting hazard information for REACH. For each data Klimisch codes need to be assigned (Klimisch et al., 1997). These Klimisch codes are indicators for the reliability of single data.

Four codes can be assigned:

- Klimisch 1 means a reliable data without restriction: the test result is generated in compliance with an internationally accepted testing guideline (e.g. OECD TGs, GLP and QA statements) and is fully documented;
- Klimisch 2 means a reliable data with restriction: the test result is similar to a result from a guideline study, it is well documented and scientifically valid;
- Klimisch 3 means that the reliability of the data is insufficient: the test result is not generated according to an accepted method or is not well documented;

DNEL = Derive No Effect Level, PNEC = Predicted No Effect Concentration, PBT = Persistency, Bioaccumulation and Toxicity.
Klimisch 4 means that the reliability of the data is not assignable: mostly because just the outcome of the test is presented but limited information is presented on the study design or how the outcome was derived and if there were any uncertainties.

Using these Klimisch codes the risk assessor identifies the key study, in case more than one study for the same endpoint is available. In general, the key study is the study with the highest Klimisch code and it should have Klimisch code 1 or 2. This result will be used for classification and labelling and to derive DNEL or PNEC.

This Klimisch quality assessment is based on OECD TG guideline type of information. Under REACH these Klimisch codes can be applied to (Q)SAR and in vitro type of information. When such information is used to fulfil the requirements for certain endpoints limited guidance is available on how to assign these Klimisch codes (Hulzebos et al., 2010, Hulzebos and Gerner, 2010). The application of these codes to other than guideline type of data will be presented in the IAS section below.

Risk assessment of chemicals used in Cosmetics

Besides REACH other regulatory frameworks like the Cosmetic Directive require the use of alternative testing methods. The Cosmetic Directive requires other approaches than animal testing (EC-Cosmetic-Directive, 2008 included in the update of 2009). In this regulation single dose testing is not allowed. Information on acute toxicity needs to be retrieved from other sources than from animal testing. For skin and eye irritation the animal tests were banned, though for the latter an officially validated test to distinguish eye corrosion from eye irritation is not yet available. In 2013 this Directive wants to exclude all animal testing which includes application of repeated doses to animals. Such application tests are now standard for the sensitisation endpoint, repeated dose toxicity endpoint and the reproductive toxicity endpoints. Adler et al. (2011) evaluated extensively the potential of alternatives for these endpoints and concluded that for these endpoints no valid methods will be available soon.

Because chemicals registered for REACH can also fall under the Cosmetic Directive there is a conflict between the requirements of the regulation and this directive. A report to address this has been sent to the European Parliament (EC, 2011).

Risk assessment of chemicals and new initiatives

The need for assessing the risks of chemicals also induced the need for alternative approaches for these assessments in the EU and in industrialised countries such as the US, Canada, Japan and Australia. A report of the Gezondheidsraad (2001) and a paper by Fentem et al. (2004) highlighted the need for a paradigm shift in toxicology to risk assessments that support
decisions about consumer safety without the need to generate data through animal tests. This paper by Fentem (2004) was the starting point of the Assuring Safety without Animal Testing (ASAT) initiative in 2008 in the EU (ASAT, 2008). In the new strategy in silico and in vitro testing should replace step by step in vivo testing using a risk based approach for the assessment of the safety of food, pharmaceuticals, cosmetics and chemicals.

In 2002, the US-EPA recognised the need to review existing strategies and to develop a long-range vision for toxicity testing and assessment, partly inspired by the European work. This resulted in a vision presented in the NRC report: TOX 21, Toxicity Testing in the 21st Century, a Vision and Strategy (NRC, 2007, Krewski, 2010). In this vision, recent advances in systems biology, (Q)SARs, testing in human cells and tissue (in vitro testing) as well as “omics” e.g. toxicogenomics, proteomics and metabolomics are claimed to change the way chemicals are tested for risk they may pose to humans. Toxicogenomics evaluate all possible biological responses in a target tissue at the level of gene expression; proteomics evaluate most biological responses at the level of protein expression; and metabolomics evaluates changes in the profiles of a large number of small molecules usually in clinically accessible fluid like blood plasma (ECETOC, 2010). The ASAT and TOX 21 approaches look into new developments for hazard assessment. The other objective is to get more information on the key features resulting in an observable effect after exposure to a chemical. The focus of hazard identification will need to shift from outcome of the study to molecular mechanisms. One of the issues is that animal testing may not be that relevant for humans because of different metabolic and biological pathways. Another issue is that the high doses tested in animals compared to the low doses that humans and the environment are in general exposed to.

Therefore, the development, validation and integration of alternative methods to animal testing have become increasingly important in the first decade of 2000. An extensive overview of in vitro testing for regulatory purposes is presented by ECVAM (Zuang et al. 2010, Adler et al. 2011, and Voutchkova et al. 2010).

(Q)SARs

Since the evaluation of (Q)SARs for regulatory purpose plays an important role in this thesis, first some historical background will be presented before we explore the recent developments. Because the use of (Q)SARs is not undisputed the pros and cons will be discussed. Next we will provide some background information on testing strategy development and the need to build a harmonised system for data evaluation from different sources. Finally some remarks will be made on the structure of this thesis.
History of (Q)SARs

Though the first work on (Q)SARs dates back to the second half of the 19th century, the founders of the modern (Q)SARs are Hansch and Hammett (Fig. 2). In the thirties Hammett correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity and some twenty years later Hansch recognized the importance of including the hydrophobicity parameter, expressed as the octanol-water partition coefficient (Log Kow), as an indicator for biological availability (Hansch, 2011, Bevan, 2011). The log Kow is now a well-known parameter to provide a measure of the bioavailability of chemicals, which will determine, in part, the amount of the compound that reaches to the target site. Furthermore Hansch and co-workers (1962, 1964, and 1969) initiated the development of QSARs for several disciplines in science. The application of these QSARs was implemented in the pharmaceutical and agricultural industries especially to find more active compounds (Rekker, 1992).

Fig. 2 The founders of the (Q)SAR methods, Corwin Hansch (1918-2011, left), Louis Hammett (1894-1987, right).

In the Netherlands Könemann (1981) initiated QSAR development for the hazard assessment of aquatic organisms, which was continued by Hermens and Verhaar in the nineties (Verhaar, 1995). It was shown that the acute aquatic effects of narcotic acting chemicals (with no or limited chemical reactivity or baseline toxicity) could be well predicted based on their log Kow (Könemann, 1981). This relationship formed the basis of the ecotoxicological (Q)SARs by Nabholz for the Environmental Protection Agency (US-EPA) to substantiate the presence or absence of concern for the environment of PreManufacturingNotifications (PMN) with limited data. They used the Könemann (Q)SAR to prove that a chemical was hazardous before requiring further information (Wagner, et al., 1995). The US-EPA used structural alerts to define more reactive chemical classes compared to baseline toxicity for a better prediction of these types of chemicals for aquatic toxicity. Around 150 structural alerts, called chemical classes are now available to predict the toxicity, short and long-term for algae, Daphnia (invertebrate) and fish (vertebrate) to assign a potential hazard and risk to the chemical submitted in the PMNnotifications. They established a computerised tool called ECOSAR with which EC50, LC50 and NOEC values can be calculated (US-EPA-EPISuite, 2011).
**Introduction into (Q)SARs**

In general a distinction can be made between SARs and QSARs as defined in the REACH guidance:
- a SAR is a qualitative relationship that relates a (sub)structure to the presence or absence of a property or activity of interest. The substructure may consist of adjacently bonded atoms, or an arrangement of non-bonded atoms that are collectively associated with the property or activity;
- a QSAR is a quantitative relationship (often a statistical correlation) relating one or more quantitative parameters derived from chemical structure to a quantitative measure of a property or activity (e.g. an (eco)toxicological endpoint). QSARs are quantitative models yielding a continuous or categorical result (EC, 2008, Chapter 6).

The term (Q)SAR is not used in a consistent way: in some cases, the term quantitative is used to refer to the nature of reporting the endpoint, whereas in others it refers to the nature of the parameters in the model. In this thesis we use the term (Q)SAR when the nature of the parameters, used to make the prediction, is quantitative but the prediction of the endpoint is qualitative. One example is the CAESAR model for skin sensitisation and is presented later. The term (Q)SAR is also used when both types are simultaneously addressed.

Both SARs and QSARs are built for finding optimally active pharmaceuticals or pesticides in a series of compounds with similar structures. They are often based on the fit of the basic (congeneric) chemical structure in the receptor. (Q)SARs are used to find molecular substructures to increase the activity of these chemicals. After finding a number of potentially active chemicals, these will be further tested in *in vitro* and *in vivo* test batteries to find the best acting ones. (Eco)toxicological (Q)SARs are more often built on data bases of *in vivo* test data of specific (eco)toxicological hazard endpoints.

QSARs for ecotoxicity have also been built on similarities in types of toxicity such as neutral organics (narcotic acting) which are toxic because they can disturb lipid membrane functions. QSAR can also be clustered based on reactivity profiles (Verhaar, 1995). These profiles are based on electrophilic substructures in their molecular structure e.g. containing alpha-beta conjugated bonds (see section below on “QSARs for ecotoxicity” for an example of such a structure). Such profiles have also been derived for skin sensitisation by Roberts and Aptula (e.g. Aptula and Roberts, 2006, Roberts et al., 2007a).

QSARs have been built on non-congeneric groups of chemicals also. These types of QSARs often use many chemical descriptors and statistical analyses to identify the key descriptors which identify the (eco)toxicity (Papa et al. 2011, for acute oral toxicity in mice).
In the CAESAR model on skin sensitisation statistical analysis is used to identify the nearest analogues but it only predicts the presence and absence of skin sensitisation and is therefore considered a (Q)SAR model (CAESAR, 2011).

It is important to mention that predictions of all types of (Q)SARs methods should only be used when the chemical of interest (query chemical) is in its applicability domain. It will be shown in this thesis that (Q)SAR models have different ways to present this fit in the domain and additional expert judgment is often needed.

(Q)SARs aim to predict the likelihood of the occurrence of (eco)toxicological effects after the exposure to certain chemical compounds by means of a relation of such effects to their molecular and physico-chemical properties. Such (eco)toxicological effects are the result of two distinct physiological phenomena, namely bioavailability (toxico-kinetic aspect) and interaction with a target inside the organism (toxico-dynamic aspect), also referred to as MoA.

The overall bioavailability is determined by the toxico-kinetic parameters: absorption, distribution, metabolism and excretion (ADME parameters). The interaction of the chemical with the body on cell level is determined by toxico-dynamic parameters such as disturbing or disrupting lipid membrane functions or by protein, DNA and receptor binding. This binding is followed by a biological cascade of activities, finally resulting in the phenotypic (eco)toxicological effect: mortality, erythema, or decrease of testis weight in rats (Aptula and Roberts, 2006). In Fig. 3 a schematic view of all stages in the MoA of a chemical is presented (Richard, 1995). One can distinguish the MoA from the mechanism of action. MoA is the description of key events or processes by which a chemical causes a disease state or other adverse effect. The mechanism of action is a description, often at the molecular level, of the means by which an agent causes a disease state or other adverse effect (Krewski et al., 2010, IPCS, 2007).

Most (Q)SARs in (eco)toxicology for regulatory purpose use the relation between the chemical structure and/or its descriptors and the guideline study result. In this way the (Q)SAR prediction takes into account both the toxico-kinetics and toxico-dynamic properties of the chemical. It is, however, difficult to determine the exact feature of a chemical that is responsible for the outcome of the test. For example, for skin irritation, skin sensitisation and genotoxicity these features are much better established than for systemic and reproductive toxicity.

In this thesis the SAR definition is applied to models which use structural alerts. The alerts are defined by using 2-D structures. These alerts can give a
qualitative value for the presence or absence of a toxicological activity/effect. The Derek for Windows for human health endpoints and the skin irritation

Exposure – Bioavailability - Chemical reactivity - Biological effects - Mode of Action

**Toxico-kinetics:**  
Absorption  Distribution  Metabolism  Excretion

**Toxico-dynamics:**  
Reactivity of the chemical and Receptor binding

![Schematic view on the relationship between SAR properties and the toxico-kinetic and toxico-dynamic aspects of the exposure to a chemical in the body](from Richard, 1995)

Fig. 3  

Single (Q)SARs are often built into (Q)SAR computer software programs. A query chemical can be added to the program and the program selects the (Q)SAR which is applicable for this chemical. Three of these programs, viz. Derek for Windows, TOPKAT and ECOSAR, have been used in this thesis to predict the effects for some chemicals and for a validity assessment (Hulzebos and Posthumus, 2003, Hulzebos et al., 2005b).

Derek for Windows found its origin in industry which wanted to capture the toxicological expertise of Derek, one of its much valued employees (now called Derek Nexus, LHASA, 2011). It is a SAR program showing structural alerts associated with effects on skin and eye irritation, skin sensitisation, organ
Toxicity, developmental toxicity, genotoxicity and carcinogenicity (human toxicological endpoints). The modules on skin sensitisation and genotoxicity are best developed considering the number of structural alerts to characterise the effect and the explanation of the MoA. Derek Nexus provides the number of positives and negatives in their training set. When no structural alerts are present, no prediction is presented and therefore the program is more limited in predicting negatives. The predictivity of Derek for Windows shows 50-70% success rate predictions for positives for sensitisation and genotoxicity (Hulzebos and Posthumus, 2003).

TOPKAT was developed by Enslin and is a (Q)SAR program based on the original approach of Hamnett and Hansch, where the effect is related to hydrophobicity and the electronic properties of the chemical (Accelrys, 2011). TOPKAT predicts a similar set of human health endpoints as Derek for Windows and the program also predicts aquatic toxicity. TOPKAT indicates whether the query chemical is in the applicability domain and predicts both positive and negative results. A special feature of TOPKAT is predicting a Low Observed Adverse Effect Level (LOAEL) for repeated dose toxicity. The predictivity by internal validation resulted in ca. 80% predictivity. Venkataphaty et al. (2004) claim that TOPKAT predicts the LOAEL in mg/kg body weight (bw) correctly in 93% of the chemicals: i.e. within a factor of 5 of the established LOAEL in the repeated dose toxicity test.

As mentioned before ECOSAR was developed by the US-EPA as a tool for hazard assessment for the PMNotifications. It is also a computer expert program and shows correct predictions between 70 and 90%, depending on the allowed uncertainty of a factor 10 maximally (Hulzebos and Posthumus, 2003, Tunkel et al., 2005, de Roode et al., 2006).

**Analogue and Category approaches**

The analogue and category approaches are basically SAR methods. They are based on the same premises, that chemicals with structural similarity will have similar MoAs (note that the actual filling of the data gaps is mostly referred to as read across). The (eco)toxicity of chemicals may be similar when they have:

- A common backbone and a common functional group (e.g. the latter being an acid (COOH), aldehyde, epoxide, ester, specific metal ion);
- an incremental and constant change across the category (e.g. a chain-length category, e.g. formic acid, acetic acid and propionic acid: C1-C3 and COOH);
- The likelihood of common precursors and/or breakdown products, via physical or biological processes, which results in structurally similar chemicals (e.g. the metabolic pathway approach of examining related chemicals such as acid/ester/salt). An example of such an approach is presented by Wu et al. (2010) on the terpinoid alcohols and the related esters (see Fig. 4). A number of categories have been evaluated in the

![Chemical structures](image)

Fig. 4 Terpinoid alcohols and ester can be placed into one category because via chemical and biological processes alpha-terpineol is the main metabolite (from Wu et al., 2010).

**Available (Q)SARs for human health endpoints**

For the absence or presence of skin and eye irritation SARs have been developed by Gerner, Hulzebos and Walker and co-workers. This work has been integrated in the SAR model on skin irritation in Chapter 4 and therefore some additional detail is presented. Gerner et al. found that for skin irritation most chemicals do not have to be classified and labelled based on the New Chemical’s database. The percentage of new chemicals that did not induce any kind of irritation on the rabbit skin was 54.1%, 7.1% had to be labelled for skin irritation, and 5.7% for skin corrosion. The remaining 33% showed skin irritation properties but the irritation scores that are needed for classification were not achieved (Gerner et al., 2008). This triggered Gerner and co-workers to develop rules on the absence of skin irritation and skin corrosion using physico-chemical properties of almost 1300 substances (e.g. Gerner et al., 2008). They found, for example, that all chemicals in this database with a log Kow < -3 do not show irritant or corrosive effects. Such rules were developed for several empirical classes: molecules containing only carbon, oxygen and hydrogen do not have irritating properties when their melting points are > 55 °C. Their set of rules was validated with another set of new chemicals and the prediction was correct for > 98% (Rorije and Hulzebos, 2005, Hulzebos et al., 2005b). It was found that 42.3% of the testing of new chemicals can be prevented. This set of rules has been developed at the Bundes Institut für Risikobewertung and is therefore called the BfR Rulebase.

Gerner et al. also developed physico-chemical exclusion rules for eye irritation. These are further discussed in Tsakovska et al. (2005). Worth and his co-workers have been programming the eye and skin irritation (Q)SARs into Toxtree (JRC-Toxtree, 2011).
For skin sensitisation Roberts, Aptula, Patlewicz and Mekanyan have developed chemical reactivity classes to predict the presence of skin sensitisation, based on electrophilic centres in the molecules. These classes have been included in the JRC-Toxtree and OECD-Toolbox (2011). The Times-SS (Tissue MEtabolisation for Skin Sensitisation) model uses similar and additional structural alerts to predict sensitisation and moreover it includes potentially sensitising metabolites. Advantages over the OECD-Toolbox prediction are that this model predicts active and non-active sensitisers, it uses semi-quantitative classes to address the potency of the sensitisers (Patlewicz et al., 2007), and it has been externally validated. It predicted the absence of skin sensitisation for 87.5% and the presence of skin sensitisation for 56% of the chemicals correctly, respectively (Roberts et al., 2007b).

Miller and Miller (1981a and b) established relationships between the chemical substructures (structural alerts) and carcinogenicity, which were refined by Ashby and Tennant (1988) and updated and reviewed by Benigni and Bossa (2011). This resulted in the ‘supercarcinogen’, an artificial molecule containing all the structural alerts for carcinogenicity. This genotoxic type of (Q)SAR information has now been included in all expert systems to a certain extent, such as Derek for Windows and TOPKAT. Newly developed chemicals hardly contain these alerts or the potency of these alerts is diminished. For example by incorporating sterically hindering atoms or atoms which increase the solubility and in these ways lose their reactivity (reviewed in Benigni and Bossa, 2011). This success is also due to the strict regulatory actions to exclude genotoxic chemicals.

For repeated dose toxicity only few (Q)SARs exists. Derek for Windows contains a number of SARs for liver, kidney and thyroid toxicity and haematological toxicity. TOPKAT is the only program which predicts a LOAEL and quantifies the prediction (Accelrys, 2011, OECD, 2004).

SARs and grouping approach have also been used to predict reproductive toxicity, but the number of substances that can be assessed for safety with these approaches is still very limited (DeVito, 1996; Maslankiewicz et al., 2005; Hulzebos and Posthumus, 2003; Fabjan & Hulzebos et al., 2006; Piparo and Worth, 2010; Hewitt et al., 2010 and Wu et al., 2010).

Especially in SAR approaches the absence of structural alerts in a chemical structure often means that the prediction is out of the applicability domain, which is especially important for those endpoints with limited number of alerts such as for repeated dose and reproductive toxicity.

General reviews on (Q)SARs for human toxicity have been presented by Vouthchkova et al. (2010) and Tsakovska et al. (2008). For the human health endpoint the US-EPA specified chemical classes addressing the toxicity of e.g. ethylene glycol ethers, certain metals etc. but these classes have a more
Available QSARs for ecotoxicological endpoints

In addition, to the work by the US-EPA, Verhaar and Hermens have done extensive work to classify chemicals into four major mechanistic classes (Verhaar, 1995). Chemicals may act according to following the mechanisms in increasing order of toxicity:

- neutral organic mechanism is associated with altered structure and functions of the cell membrane and shows an anaesthetic like of effect (low reaction to stimuli). Examples are alcohols (but not methanol) and chlorobenzenes, such as n-Butanol and Pentachlorobenzene.

![n-Butanol and Pentachlorobenzene](image)

- polar organic mechanism seems to result in a somewhat higher toxicity compared to neutral organics especially for chemicals with a lower Log Kow (< 2.9). It is not clear why this difference exists, it may be due to some ionisation or additional hydrogen bonding. A third reason may be that neutral organics partition themselves in a 3-D manner while polar narcotics to that in a 2-D way (Worth et. al., 2007, in van Leeuwen and Vermeire, Chapter 10). This type includes the chlorophenols and 2,4,6-trichlorophenol is presented as an example.

![2,4,6-Trichlorophenol](image)

- Reactive mechanisms generate a higher toxicity compared to the neutral and polar organics and can be defined by structural alerts. The reactivity can be well explained in terms of reactivity to proteins due to electrophilic (positively charged) or nucleophilic centres (negatively charged) in the molecule. For example esters are more reactive compared to neutral and polar organics. Their reactivity with proteins is explained by electron-withdrawing atoms such as chloride attached close to the ester bond (see Fig. 7). In this structure a second electrophilic centre is shown on the left side of the molecule, which is a 1,3-conjugated diene.
In addition, Available Q (Q)SARs show that they can predict the likelihood of the presence or absence of (eco)toxicological effects and thus limit animal testing and costs.

Screening chemicals with (Q)SARs prevent marketing of chemicals with toxic effects. Before bringing the chemical to the market adjustments can be made to the structure to have similar functionality but diminished toxicity. For example, toluene is much less toxic than benzene with one additional methyl group on the benzene ring. Several of such examples are presented by DeVito (1996). The
success of the elimination of structural fragments to diminish genotoxicity has already been presented in the section above where the work of the Millers is discussed.

Besides for screening purposes (Q)SARs often help to determine whether the MoA of chemicals is similar. The sensitisation alerts defined by Aptula and Roberts (2006) are grouped into similar chemical mechanisms such as Michael addition, Schiff base, SnAR, Sn2 reactions, which all indicate qualitatively the skin sensitisation potential. Within these reaction mechanisms a more quantitative approach may also be feasible. Similar reactivity groups have been distinguished within ecotoxicity. The distinction of narcotic and polar narcotics mechanisms was based on the different slopes of the regression lines for these two groups (Worth et al., 2007 in van Leeuwen and Vermeire, Chapter 10).

Though these available (eco)toxicological examples show that there is certainly a place for (Q)SARs in hazard assessment, this possibility is not embraced by everyone. Especially, at the beginning of 2000 when (Q)SARs were promoted as alternatives for animal testing, many scientists claimed that (Q)SARs can only be used in a limited number of cases for hazard assessment purposes. First of all there was a lack of criteria and guidance to assess the validity of the (Q)SARs, especially for those concerning human toxicological endpoints. In addition, no appropriate and validated (Q)SAR-tool/systems were lacking. They also stated that the applicability domains, predicting power and transparency of (Q)SARs, are often diffuse. Also the complexity of the endpoint and the high or unknown variability of the underlying test data are often ignored (Hulzebos and Posthumus, 2003).

Moreover these critics do not believe that (Q)SARs can adequately predict the complex route and effects of chemicals. They question whether the bioavailability of the chemical has been taken into account; has the reactivity with a protein or DNA been covered; does the (Q)SAR include the metabolisation and does it predict the biological cascade of processes finally leading to the observed MoA (see Fig. 3, Richard, 1995, Aptula and Roberts, 2006). Though several non-testing (e.g. analogue approaches) and testing methods (in vitro and in vivo) besides (Q)SARs are available to predict the toxicity of chemicals, there are several examples, showing that the toxicity of some pharmaceuticals and pesticides was not adequately predicted by the safety screening methods used, including (Q)SARs (e.g. Olson et al., 2000). This may be due because of sensitivity differences within and between species, different metabolic pathways in species and unexpected metabolic routes.

There are several examples of chemical outliers in a series of similar structures. Softenon, causing severe malformations of the limbs of new born, is a well-known example that one enantiomer of the chemical is causing these effects while the other enantiomer does not. However, it is unfair to blame the Softenon disaster on (Q)SARs, since this was not the only method used in the
screening process. The alkanes are mentioned earlier as chemicals with low toxicity (Hau et al., 2002). However, n-hexane has a higher toxicity than would be expected from its structure and the neighbouring alkanes. This is due to the formation of a toxic metabolite: 2,5-hexanedione (DeVito, 1996, Rila et al., 2006). In 2010 such toxicity was seen in Chinese workers using n-hexane illegally (People Daily online, 2011). Also methanol is more toxic compared to longer chain alcohols, partly due to the formation of formaldehyde as a metabolite. This may also be the reason that methyl esters with methanol as a hydrolysis product are sometimes more toxic compared to their higher chain length analogues e.g. methylcarbamates (OECD, 2007). This means that when evaluating chemicals using alternative methods not only the parent chemical has to be evaluated but also the potential metabolites.

(Q)SAR can elucidate these differences in toxicity by defining the structural alerts, chemical descriptors and their applicability domains. Organophosphate activity, for example, is limited to those chemicals which have relatively short C-chains attached to the phosphate. Chain lengths > C4 considerably limit the activity of the pesticide (LHASA-Derek Nexus, 2011).

(Q)SARs and application for regulatory hazard assessment of chemicals

In spite of the above mentioned criticisms further research into QSARs is stimulated because they can be consulted as a first or additional safeguard. In this thesis it is shown that its use may lead to limited animal testing, which nowadays has high priority in many levels of society (EC, 2006, NRC, 2007, ASAT, 2008, EC-Cosmetic Directive, 2009). To make (Q)SARs more applicable for regulatory purposes the following principles should be addressed and documented:

1) the defined endpoint for a regulatory purpose;
2) the method or algorithm;
3) the applicability domain (or width) of the method;
4) the predictivity of the model and;
5) a mechanistic interpretation, if possible.

These OECD principles on (Q)SARs were the start of the coordinated assessment and documentation of existing (Q)SARs along these criteria (OECD, 2004 and 2007). The OECD published a working document for several expert systems such as Derek for Windows, TOPKAT and some of the EpiSuite models (OECD, 2004). ECETOC has presented evaluations on commercial QSAR programs (ECETOC, 2003).

The former European Chemicals Bureau (moved partly into ECVAM (being part of the Institute for Health and Consumer Protection (IHPC)) and partly to ECHA in 2008) played a central role in this process. They organise workshops, by establishing a platform for (Q)SAR tools and by doing several (Q)SAR evaluations themselves. Quite a variety of models are presented on the website
of the IHPC including those for eye and skin irritation from Gerner et al., (e.g. 2008) which are both based on New Chemical information from industry (JRC-Computational Toxicology, 2011). For further support of using (Q)SARs for regulatory purpose, ECB funded several projects to apply the OECD principles on (Q)SAR to existing and newly developed models which are reviewed in Worth et al. (2007). The validity of non-testing approaches such as (Q)SARs and grouping approaches have been carried out for skin irritation/corrosion, skin sensitisation and genotoxicity and reproductive toxicity (Rorije and Hulzebos, 2005, Roberts et al., 2007c, Benigni et al., 2007). Also other publications verified models for their compliance with the OECD principles as has been done by Fabjan and Hulzebos (2008) for sensitisation alerts. These publications show that the OECD principles for the validation of (Q)SARs have led to practical criteria for assessing these. The application of these criteria into QSAR Model Reporting Format (QMRF) and QSAR Prediction Reporting Formats (QPRF) for each model and each prediction facilitates the REACH application of (Q)SARs (JRC, 2008a and b).

REACH gives more room for non-guideline methods, such as (Q)SARs compared to the former legislation on industrial chemicals. In the REACH legislation the framework of using alternative methods for assessing the hazard of chemicals is laid out in Annex XI for (Q)SARs. In this Annex it is specified that:

(i) a (Q)SAR prediction must be derived from a model whose scientific validity has been established (that is fulfilling the 5 principles of the validation of (Q)SARs);

(ii) the substance, which is predicted, must fall within the applicability domain of the model;

(iii) the results are adequate for the purpose of classification and labelling and/or risk assessment (which means that for most (eco)toxicological endpoints a quantitative value is needed) and;

(iv) adequate and reliable documentation of the applied method is provided (the QSAR model and the QSAR prediction needs to be documented).

In the REACH guidance documents there is a full chapter on (Q)SARs, explaining how these Annex XI criteria for (Q)SARs can be fulfilled (EC, 2008, Chapter 6).

**Integrated Testing Strategies (ITS)**

In the process of applying (Q)SARs for hazard assessment for regulatory purposes it became clear that for many types of chemicals single (Q)SAR predictions would not be sufficient to assess their toxicity. It was therefore emphasised that both non-testing (e.g. (Q)SARs) and testing information should be combined in testing strategies. Though testing strategies are not novel and
always have been used in the risk assessment of chemicals (e.g. Dejongh et al., 1999), under REACH it became a more a specific tool to limit animal testing.

In the REACH guidance documents tiered ITSs are provided to make an adequate decision for each required physical-chemical, environmental fate and (eco)toxicological endpoint (EC, 2006, 2008). At every endpoint these strategies are presented which take a stepwise approach for the gathering of non-testing and testing information, retrieving (Q)SAR prediction, and performing in vitro testing, before doing further animal testing. In these testing strategies assigning Weight of Evidence (WoE) by expert judgment is a key activity to finalise a decision on the (eco)toxicological endpoint and to determine whether further information is needed.

Besides the testing strategies in REACH many others have been developed in the last 10 years in order to define a decision for a particular (REACH) framework (EC, 2008, Ahlers et al., 2008, Balls, 2010, Cronin et al., 2008, 2009, Grindon et al., 2008, Vonk et al., 2009). For acute oral toxicity TNO has developed a tiered testing strategy using structural alerts for reactivity, based on the structural alerts defined by Hulzebos et al. (2005b), and on an in vitro cytotoxicity database to determine acute oral classification of substances (Freidig, et al., 2007). A tiered testing strategy is needed for eye irritation, because only validated in vitro eye corrosion tests are available (e.g. Isolated Chicken Eye test, OECD TG 438, 2009). This means that when a chemical is not corrosive in such a test still an in vivo eye irritation test needs to be done to assess the eye irritation potential. Another option could be to add a prediction of the model of Worth and Cronin (2003), which predicts the absence of eye irritation, based on molecular weight. The prediction could be sufficient for waiving the in vivo eye irritation test, possibly with additional reasoning on the low reactivity of the chemical when applicable and pointing to a chemical in the training set that is similar to the query chemical. A tiered testing strategy for skin sensitisation has been built by e.g. Natsch et al. (2010) including protein assays to simulate the covalent binding to the protein, which is one of the key mechanistic features of the skin sensitisation.

This thesis presents (Q)SARs and tiered testing strategies for skin irritation. These have been developed with main contributions from Walker and Gerner and will be discussed in Chapter 3 (Walker et al., 2005, Gerner et al., 2004). This testing strategy has been included in the Skin Irritation and Corrosion Rules Estimation Tool (SICRETool), for the evaluation of the absence and presence of skin irritation classification and labelling, and is presented in Walker et al. (2005 and in Fig. 9). This tool is one of the few established tiered testing strategies mentioned in the REACH guidance with clear criteria for the prediction of the presence and absence of skin irritation. The tool is incorporated in JRC-Toxtree and the OECD-Toolbox (2011).
Integrated Testing Strategy for Skin irritation

Evaluation of all existing toxicological information including human experience → information sufficient for assessment → Classification and labelling

Evaluation of information on:
- purity of the chemical
- CAS RN
- structural formula
- empirical formula

Information not sufficient for assessment → Evaluation of additional physico-chemical information, e.g.:
- aqueous or lipid solubility
- surface tension
- vapour pressure

Measurement or calculation of:
- molecular weight
- melting point
- partition coefficient (logKow or logPow)

Are the chemical’s physico-chemical properties outside the limit values for skin corrosion?

Are the chemical’s physico-chemical properties outside the limit values for skin irritation?

Are structural alerts for skin corrosion in the chemical’s structure?

Are structural alerts for skin irritation in the chemical’s structure?

No skin corrosion potential

No skin irritation potential

Corrosive to skin

Irritating to skin

Positive in vitro testing for skin irritation

Positive in vitro testing for skin corrosion

Data used for development and validation of structural alerts

Data used for development and validation of physico-chemical limit values

Data used for development and validation of structural alerts

Fig. 9 A tiered testing strategy to assess the skin corrosion or skin irritation properties of chemicals (Walker et al., 2005)
According to REACH, the outcome of (Q)SARs and testing strategies assessing the hazards of chemicals should be sufficient to perform the:

1) hazard (or effect) assessment;
2) classification and labelling, leading to the application of labels (pictures) and hazard phrases (the warning sentences) (See below for an example);
3) assessment of PBT-properties (Persistent, Bioaccumulating and Toxic);
4) risk characterisation in which No Effect Levels or No Effect Concentrations are compared to the exposure estimation for workers, consumers and environment, to determine if there is a safety concern.

Testing strategies may also include the assessment of the potential of exposure to the chemical. In case exposure is not significant for workers, consumers or the environment in all stages of the life cycle of the chemical further testing may be omitted as is laid out in Annex XI (3) of the REACH regulation (EC, 2006, and amendment in 2010). Testing strategies for exposure based waiving and for defining the Threshold of Toxicological Concern (TTC) can further reduce animal testing as have been discussed by Vermeire et al. (2010). When the exposure can be considered sufficiently low no risk is anticipated. This is, for example, the case when the exposure is calculated below the DNELs and PNECs. The TTC approach is based on the premise that there is a level of exposure to chemicals which would not entail a risk for human health even for genotoxic carcinogens. This level has been established at 0.15µg/person/day. Higher levels have been established for non-genotoxic types of chemicals. These latter methods are beyond the scope of this thesis.

Many tiered testing strategies are based on a step by step approach. This means that each alternative result is evaluated on its merits. When it does not fulfil the REACH required endpoint as a standalone the result is set aside. In such a process useful information may be left out. Therefore Jaworska et al. (2010) advocates ITSs, which maximises the use of existing data and gains a more comprehensive and mechanistic base for decision making using (Q)SARs, in vitro and in vivo data. They promote assigning weights to the available information as to the predictivity of the MoA information and have presented an example of this in Jaworska and Hoffmann (2010).

Though testing strategies are advocated next to other alternatives for animal testing the acceptance by the scientific community and by regulatory authorities will need a thorough justification of their appropriateness for a given purpose
(Lillenblum et al., 2008, Schaafsma et al., 2009, ECHA, 2011b). According to ECHA “an evidence based approach involves an assessment of the relative values/weight of different pieces of information that have been retrieved. To this end a value needs to be assigned to each piece of information. These weights can be assigned in an objective way using a formalised procedure or by using expert judgement. The weight given to the available evidence will be influenced by features such as quality of the data, the consistency of the results, the nature and severity of the effects, the relevance of information for the given regulatory endpoint. One definition is: the process of considering the strengths and weaknesses of various pieces of information in reaching and supporting a conclusion concerning the property of the substance” (ECHA, 2011). Therefore the use of WoE is indispensable in testing strategies for single data which by themselves are not completely fulfilling the (eco)toxicological endpoint. However, when all weighted data are combined a specific REACH decision on an (eco)toxicological endpoint can be made.

One of the main difficulties in applying a WoE is the fact that data may come from widely different methods (in silico, in vitro and in vivo), which have different uncertainties and cover different MoAs. In this thesis an objective way using a formalised approach (as mentioned in the ECHA guidance) is presented to weigh data before they enter the ITS. This process is presented in the Integrated Assessment Scheme (IAS).

**Integrated assessment Scheme (IAS)**

When combining data originating from different methods (non-animal and animal), in a testing strategy it becomes clear that there is no frame to evaluate these data in a similar way for a specific context. Therefore an important contribution of this thesis for limiting animal testing is the development of an evaluation scheme, the IAS, to assess the quality of all types of toxicity data following a similar set of criteria for a specific effect before these data can be used in a testing strategy. In assessing the quality of data for regulatory purposes three modules can be distinguished, which evaluate:

1) the reliability of the data;
2) the validity of the method, where the data are obtained;
3) the purpose of use of the information.

For each of these modules, five criteria (key principles) have been developed to answer the question on a) what is tested; b) how has this been tested; c) is the chemical suitable for this type of test; d) what is the uncertainty of the testing and is the uncertainty sufficiently small to protect human health and environment and; e) what is causing the effect. These five subjects are specifically phrased for each module to be able to evaluate these modules separately. It is clear that these five principles are very similar to the OECD principles for the validation of (Q)SARs. One of the goals of this thesis is to adapt these (Q)SAR principles in such a way that they can be used for other methods than (Q)SARs and for varying regulatory purposes. The three modules
and five key principles result in a 3 x 5 matrix. For each module assessment codes are assigned which are derived from the Klimisch codes. With these combined codes the risk assessor can decide whether information is i) adequate (standalone); ii) partly adequate (ITS is needed) or: iii) inadequate (should not be used) (Hulzebos et al. 2010, Hulzebos and Gerner, 2010).

The weighted and evaluated data can then be combined and integrated in a testing strategy to see whether the information available fulfils the purpose of the evaluation. When insufficient data are present for a particular purpose the gathering of further information is needed which may include animal and non-animal testing. The IAS is an efficient approach to bring to light the strengths and limitations of data from different methods in a standardised way (Hulzebos et al., 2010). The expert judgment indispensably needed for weighting and evaluation of data can now be transparently documented.

The use and practicalities of the IAS are illustrated in an ITS for skin irritation and corrosion using using five chemicals (Hulzebos and Gerner, 2010). This ITS had been developed earlier and implemented in the SICRETool (Walker et al., 2005). In this ITS the chemical structure features, data from the standard OECD TG and the in vitro skin irritation information is combined in an ITS using objective and formalised criteria. The aim here is to reach a decision on the presence or absence of skin irritation for a particular (regulatory) purpose, without the need for further testing.

The structure of this thesis

The predictions of the (Q)SAR tools are assessed in Chapter 2. For human health endpoints Derek for Windows and for aquatic toxicity ECOSAR are selected. The predicted values are compared with peer reviewed experimental data (Hulzebos and Posthumus, 2003).

The same (Q)SAR tools are evaluated against the OECD principles on the validation of (Q)SARs (OECD, 2004, Hulzebos et al., 2005a) in Chapter 3.

In Chapter 4 a hybrid skin irritation SAR model is developed. The model is based on literature derived structural alerts for reactivity, which are combined with physico-chemical properties indicating limited bioavailability. This model makes most skin irritation tests superfluous (Hulzebos et al., 2005b).

In Chapter 5 a novel tool, the Integrated Assessment Scheme (IAS) is presented. This tool is expected to guide the risk assessor in the evaluation of information from (Q)SARs and other (eco)toxicological information, using a similar denominator. It aims to help the risk assessor in the decision making, which information to use for a particular regulatory purpose. It can also help in the assessment whether information can be applied as standalone, an information element in a testing strategy, or should not be used (Hulzebos et al., 2010).
In Chapter 6 this IAS is illustrated using predictions of five chemicals from three (Q)SAR models: BfR Rulebase, Derek for Windows and TOPKAT. This IAS will also be applied to in vivo and in vitro information on skin irritation to show that also these types of data can be evaluated using this tool (Hulzebos and Gerner, 2010).

In the discussion, in Chapter 7, the above (Q)SAR models and the IAS will be evaluated according to the regulatory criteria defined in Annex XI of the REACH legislation where the criteria are established for using (Q)SAR predictions for REACH registrations (EC, 2006). The percentage of animal tests that can be replaced by using (Q)SAR information is estimated. In addition, the role of the IAS for further replacement is discussed.

References


Balls, M., 2010, Integrated Testing Strategies and the prediction of toxic hazard,
Issues in Toxicol., 7, 584-605.
European Community, L154.
EC, 1993, Existing Chemicals; EC Regulation 793/93, of 23 March 1993 on the evaluation and control or risks of existing substances, Off. J. Eur. Communities, L84/1.
ECETOC Technical Report, 74, 1998, QSARs in the Assessment of the Environmental Fate and Effects of Chemicals, Brussels


Gezondheidsraad, 2001, Toxicity testing: a more efficient approach (Onderzoek gezondheidsrisico’s van stoffen, een gerichtere benadering), Den Haag, in Dutch, http://www.gezondheidsraad.nl/sites/default/files/01@24N.PDF, site visited, September 2011.


Hartung, T., 2006, ECVAM's progress in implementing the 3Rs in Europe, ECVAM European Centre for the Validation of Alternative Methods, Ispra, Italy, ATLA, 23, Suppl, 21-8.


Hulzebos, E., Walker, J.D., Gerner, I. and Schlegel, K., 2005b, Use of structural alerts to develop rules for identifying chemical substances with skin irritation or skin corrosion potential. QSAR Comb. Sci., 24, 332-342.


ICCVAM, 2011, Interagency Coordinating Committee on the validation of Alternative methods, NICEATM and ICCVAM Test Method


Jaworska, J., and Hoffmann, S., 2010, Integrated Testing Strategies (ITS) – Opportunities to better use existing data and guide future testing in toxicology, Altex, 27, 231-242:


JRC, 2011, Joint Research Centre, Computational toxicology,

JRC-Toxtree, 2011, Joint Research Centre,


Natsch, A., Enter, R., and Ellis, G., 2009, Filling the concept with data: Integration data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing, Toxicol. Sci., 107, 106-121.


OECD Test Guidelines for the testing of chemicals, 2011, http://www.oecd.org/document/7/0,3343,en_2649_34377_37051368_1_1_1_1,00.html, site visited August, 2011.


OECD-(Q)SAR application toolbox, 2011, http://www.oecd.org/document/54/0,3343,en_2649_34379_42923638_1_1_1_1,00.html, site visited August, 2011.


