

Evaluation of *in silico* models for the identification of respiratory sensitizers

Running title:

SARs for respiratory sensitization

Sander Dik,^{*,†} Janine Ezendam,^{*} Albert R. Cunningham,^{‡,§,¶} C. Alex Carrasquer,[‡] Henk van Loveren,^{*,†} and Emiel Rorije¹

^{*}Centre for Health Protection and [†]Centre for Safety of Substances and Products, National Institute for Public Health and the Environment, 3720 BA Bilthoven, The Netherlands;

[†]Department of Toxicogenomics, Maastricht University, 6200 MD Maastricht, The Netherlands; [‡]James Graham Brown Cancer Center and [§]Department of Medicine and [¶]Department of Pharmacology and Toxicology, University of Louisville, Louisville, KY 40202, USA

¹To whom correspondence should be addressed at Centre for Safety of Substances and Products, National Institute for Public Health and the Environment, P.O. Box 1, 3720 BA Bilthoven, The Netherlands. Fax: +31 30 2744401. E-mail: emiel.rorije@rivm.nl

ABSTRACT

Low molecular weight (LMW) respiratory sensitizers can cause occupational asthma but due to a lack of adequate test methods, prospective identification of respiratory sensitizers is currently not possible. This paper presents the evaluation of Structure-Activity Relationship models (SARs) as potential methods to prospectively conclude on the sensitization potential of LMW chemicals. The predictive performance of the SARs calculated from their training sets was compared to their performance on a dataset of newly identified respiratory sensitizers and non-sensitizers, derived from literature. The predictivity of the available SARs for new substances was markedly lower than their published predictive performance. For that reason, no single SAR model can be considered sufficiently reliable to conclude on potential LMW respiratory sensitization properties of a substance. The individual applicability domains of the models were analyzed for adequacies and deficiencies. Based on these findings, a tiered prediction approach is subsequently proposed. This approach combines the two SARs with the highest positive and negative predictivity taking into account model specific chemical applicability domain issues. The tiered approach provided reliable predictions for one third of the respiratory sensitizers and non-sensitizers of the external validation set compiled by us. For these chemicals, a positive predictive value of 96% and a negative predictive value of 89% was obtained. The tiered approach was not able to predict the other two thirds of the chemicals, meaning that additional information is required and that there is an urgent need for other test methods, e.g. *in chemico* or *in vitro*, to reach a reliable conclusion.

Keywords:

Chemical allergy, Respiratory sensitization, (Q)SAR, Alternatives to animal testing, Predictive toxicology, Tiered approach

INTRODUCTION

Occupational asthma (OA) is a form of respiratory allergy that is characterized by bronchial constriction and bronchial hyperresponsiveness after inhalation of a sensitizing compound at the workplace. These compounds are distinguished into high molecular weight (HMW) or low molecular weight (LMW) respiratory sensitizers (Pralong *et al.*, 2012; Sastre *et al.*, 2003). In many industrialized countries, OA is a major problem because it has serious medical and socioeconomic consequences (Dewitte *et al.*, 1994) and is implicated in 10 to 17% of the total adult asthma cases (Mapp *et al.*, 2005; Toren and Blanc, 2009). To date, more than 400 substances are known to cause OA, of which a substantial part are LMW respiratory sensitizers (CSST, 2013; Mapp, *et al.*, 2005).

Over the past years, several new LMW respiratory sensitizers have been identified based on human case studies showing that the chemical caused OA (Bloemen *et al.*, 2009; Pralong, *et al.*, 2012; Quirce and Sastre, 2011). This illustrates the importance of prospective identification of respiratory sensitizers to devise protective measures for workers. Many occupational asthmagens are also respiratory sensitizers, but not all asthmagens induce immunological mediated hypersensitivity. An occupational asthmagen might also induce non-immunological hypersensitivity such as an irritant effect on the lung. This difference is difficult to distinguish in a clinical setting (Burge *et al.*, 2012), so determining whether an occupational asthmagen is a true respiratory sensitizer is challenging.

In several chemical legislations, including the European Regulation on Classification, Labeling and Packaging of substances and mixtures (CLP) (EU, 2008), REACH (Registration, Evaluation and Authorization of Chemicals) (EU, 2012), as well as the Globally Harmonized System on classification and labeling of chemicals (GHS) (UN, 2007), respiratory sensitization is considered an important toxicological end-point. REACH specifically provides guidance regarding testing for respiratory sensitization. However, the

guidance is still limited and concludes that there is a lack of widely accepted means of testing for respiratory sensitization. Some predictive animal (Arts and Kuper, 2007) and non-animal (Boverhof *et al.*, 2008; Lalko *et al.*, 2012) test methods for the identification of respiratory sensitizers have been described in the literature, but these are not widely accepted yet, nor close to the point where they could enter formal validation.

In recent years, much progress has been made regarding animal-free predictive tests methods for several toxicological endpoints, for example skin sensitization (Vandebriel and van Loveren, 2010). More recently, some of these skin sensitization methods have been adapted in such a way that they might be able to identify respiratory sensitizers. Among these methods are cell-based methods (Boverhof, *et al.*, 2008), an *in chemico* method measuring peptide reactivity (Lalko, *et al.*, 2012), and several *in silico* approaches (Structure-Activity Relationships (SARs)). The latter are computer-based models or alerts that predict respiratory sensitization based on theoretical properties such as electrophilicity (Enoch *et al.*, 2009), mechanism of reactivity (Enoch *et al.*, 2010) and/or the presence of specific chemical substructures (Cunningham *et al.*, 2005; Graham *et al.*, 1997; Jarvis *et al.*, 2005; Sanderson and Earnshaw, 1991). These SARs have been reported to identify respiratory sensitizers and respiratory non-sensitizers with accuracies varying from 69% to 94%.

None of the available SAR models for respiratory sensitization are generally accepted for regulatory purposes. Nevertheless, the models, based on their published performances, seem to hold promise for providing a fast and economic way to identify the respiratory sensitization potential of new substances. The aim of the current study was to evaluate the predictive performance of the SARs using a set of respiratory sensitizers and non-sensitizers that were new to the existing SAR models. Additionally, it was explored whether combining the best performing SARs in a tiered approach has added value in terms of predictivity.

MATERIALS AND METHODS

SAR models for predicting respiratory sensitization

A literature search was conducted to find SAR models developed for the prediction of respiratory sensitization potential. Both PubMed and Google Scholar were searched for publications between 1990 and 2013 using the following terms: computational, model, respiratory sensitization, SAR, QSAR. Based on title and abstract, three SARs were found; Graham using MultiCASE software (Graham, *et al.*, 1997), Cunningham using cat-SAR (Cunningham, *et al.*, 2005) and Jarvis using a logistic regression model (Jarvis, *et al.*, 2005). These models identify substructures of the chemicals in the respective model training sets, associated with the potential of the chemicals to cause respiratory sensitization. A new chemical is analyzed for the presence of fragments identified from the training set. The presence of activating and inactivating fragments is then used in a computer-optimized algorithm to generate a prediction for the chemical. In addition to these three SAR models, two sets of structural alerts to identify respiratory sensitizers were evaluated: the combined respiratory sensitization and occupational asthma knowledge databases of the Derek Nexus expert system (Lhasa Ltd., 2014) and a set of alerts proposed in a paper by Enoch *et al.* (2012). For this paper, the alert databases are applied as SAR models i.e. they are also used to predict the absence of respiratory sensitization potential. These five models are shortly described below.

- (i) Graham *et al.* (1997) used 40 respiratory sensitizers identified from case studies and 40 respiratory non-sensitizers selected on their negative human skin sensitization data to develop a model using the MultiCASE software (Klopman, 1992). An updated implementation of this MultiCASE model to reflect new insights in structure information was developed by the Danish Environmental Protection Agency (DK EPA) to generate information on respiratory sensitization potential. Predictions

from the MultiCASE model were obtained from the model developers directly. Several, but not all, predictions are also publically available in an online database (*Danish (Q)SAR database*). There are no original published data available for the predictive performance of MultiCASE based on the 80 chemicals from Graham, the MultiCASE implementation training set was obtained through personal communication.

- (ii) The cat-SAR model (Cunningham, *et al.*, 2005) used the same training set as Graham *et al.* (1997), but applies a different algorithm. Predictions can be obtained from the model developers directly.
- (iii) Jarvis *et al.* (2005) designed a SAR model based on 78 respiratory sensitizers from case studies and a set of 301 respiratory non-sensitizers based on occupational exposure limits. The model is available as a web-based interface (Jarvis, *et al.*, 2005).
- (iv) The Derek Nexus knowledge base (Lhasa Ltd., 2014) (from this point on referred to as Derek) contains data from published sources. Derek is continuously updated using new data; hence, for the purpose of this study, the training set of Derek is defined as all chemicals that have contributed to defining the respiratory sensitization alerts up to December 2013.
- (v) Enoch *et al.* (2012) proposed 52 mechanism-based alerts for respiratory sensitization, all based on an analysis of organic chemistry related to the molecular initiating event of covalent binding to proteins. Their training set contained 104 respiratory sensitizers and 82 respiratory non-sensitizers using the same criteria for chemical selection as Jarvis *et al.* (2005).

Compilation of a set of identified respiratory (non-)sensitizers

To evaluate the performance of the SARs, a dataset of organic LMW chemicals was compiled using the original training sets of the individual SARs and complementary literature sources (Bloemen, *et al.*, 2009; Quirce and Bernstein, 2011). Only organic molecules with a molecular weight lower than 1 kD were selected for this study. All sources considered chemicals as respiratory sensitizers or occupational asthmagens either based on criteria of the US Department of Health and Human Services “Guidelines for Diagnosis and Treatment of Asthma” (USDHHS, 1991), or because a case of OA was reported in a peer-reviewed report (Enoch, *et al.*, 2012; Jarvis, *et al.*, 2005). An additional 17 respiratory sensitizers were identified from the complementary literature sources. For the purpose of this paper, it was assumed that all occupational asthmagens are also respiratory sensitizers.

The criteria defining respiratory non-sensitizers also varied between the SAR models. A chemical was considered a respiratory non-sensitizer when it was either present in the 1994 edition of the UK Health & Safety Executive’s document *Occupational exposure limits* and no cases of OA were reported in the UK (Enoch, *et al.*, 2012; Jarvis, *et al.*, 2005; and references therein, see supplementary table 1), or if a substance tested negative for human allergic contact skin sensitizing ability (Graham, *et al.*, 1997). The first criterion is based on the assumptions that (i) the presence of an occupational exposure limit is evidence that humans have had fairly extensive industrial inhalational exposure, and (ii) a lack of reports on OA due to exposure to these chemicals therefore suggests a lack of respiratory sensitizing potential (Enoch, *et al.*, 2012; Jarvis, *et al.*, 2005). The second criterion is based on the mechanistic assumption that human skin non-sensitizers will also not be able to induce respiratory sensitization (Cunningham, *et al.*, 2005; Graham, *et al.*, 1997).

The set of respiratory non-sensitizers was subsequently expanded by adding 168 chemicals

that tested negative in the murine Local Lymph Node Assay (LLNA) (*NICEATM LLNA database*, 2013; individual references therein), that were not already found in other literature sources and for which structure information was available. These chemicals were included as they are assumed to be lacking the potential to cause respiratory sensitization, as previously suggested among others by Dearman et al. (2013). In total, a dataset of 138 respiratory sensitizers and 521 respiratory non-sensitizers was compiled from all sources to evaluate the predictive performance of the selected models. The list of all chemicals, including their individual references, is given in supplementary table 1.

Model predictions

The output of the five SARs is either numerical or textual. MultiCASE predictions are summarized as: POS(itive), NEG(ative) or EQU(ivocal). MultiCASE also applies a strict applicability domain (AD) call as determined by an AD algorithm from the DK EPA [personal communication]. MultiCASE out-of-domain and equivocal predictions were excluded from our evaluation. The model of Jarvis and the cat-SAR model generate a numerical value between zero and one and both models use a cut-off value of 0.5 to distinguish predicted respiratory sensitizers from non-sensitizers. Derek generates a textual prediction (certain, plausible, probable, equivocal, doubted, improbable, or nothing to report) based on the presence or absence of alerts and the certainty of the evidence used to define the alert. Compounds predicted as ‘certain’, ‘plausible’ or ‘probable’ were considered predicted respiratory sensitizers. Predictions of ‘equivocal’ were excluded from our analysis. All other outcomes were interpreted as the chemical being predicted a respiratory non-sensitizer. The 52 alerts from Enoch were coded into a manually operable software implementation using Accord for Excel (*Accelrys*, 2010). The presence of any of the 52 alerts was used as an indication that a chemical is predicted to be a respiratory sensitizer, and the absence of all 52

alerts was interpreted as the chemical being predicted a respiratory non-sensitizer. Both Derek and Enoch do not support the interpretation that the absence of any alert indicates that a chemical is a respiratory non-sensitizer, but this was nevertheless used to evaluate the ability of Derek and Enoch to predict the absence of respiratory sensitization potential.

All predictions from all models, as provided in supplementary table 1, were obtained by either applying the model manually or by personal communication with the respective model developers.

Characterization of the predictive performance of respiratory sensitization models

The predictive performances of the models were characterized using Cooper statistics (Cooper *et al.*, 1979); sensitivity (correctly predicted respiratory sensitizers as a fraction of the total number of true respiratory sensitizers), specificity (correctly predicted respiratory non-sensitizers as a fraction of the total number of true respiratory non-sensitizers), positive predictive value (PPV, correctly predicted respiratory sensitizers as a fraction of all positive predictions) and negative predictive value (NPV, correctly predicted respiratory non-sensitizers as a fraction of all negative predictions). Sensitivity and specificity are an indication of the diagnostic performance of a model (Doubilet, 1983); PPV and NPV are a measure of the reliability of a prediction from a model. PPV and NPV are however dependent on the ratio of (true) respiratory sensitizers and non-sensitizers in a dataset, as opposed to sensitivity and specificity. To compare the PPV and NPV derived from different testing datasets, the datasets were scaled to an equal ratio of sensitizers and non-sensitizers. A formal description of these statistics, specifically the use of scaled PPV and NPV, can be found in e.g. (Rorije *et al.*, 2013). The Jarvis and MultiCASE models were not always able to provide a prediction because these models did not accept all structure information. Furthermore, in the case of cat-SAR, when no fragments associated with activity or inactivity were identified in a

test compound, no prediction was made (i.e., cat-SAR does not make default predictions based on lack of evidence). These missing predictions were not taken into account when calculating the statistics.

The performance of each model is visualized in Receiver Operating Characteristic (ROC) plots (Soreide, 2009). These plots illustrate the performance of a model in terms of True Positive Rate (sensitivity) and True Negative Rate (1-specificity), or alternatively, in terms of PPV and 1-NPV. The upper left corner position represents perfect predictive performance. A point on the diagonal represents predictive performance similar to flipping a coin, and the lower right corner position represents a perfect, but inverted, predictive performance. Points above the diagonal therefore represent models that perform better than random. A larger orthogonal distance from the diagonal indicates better predictive performance. This allows for easy visual comparison of the predictive performance of the models.

Evaluation of the predictive performance

In order to compare the five models, different subsets from the compiled dataset were defined and for each set Cooper statistics were calculated (Table 1). The definitions of the different subsets and their abbreviated names are:

- DD: the Development Dataset is specific for each individual model, and consists of the training set chemicals from each model as published (Jarvis, Cunningham), as obtained through personal communication (Jarvis, MultiCASE) or in the case of Derek as obtained from the incorporated knowledge editor in the Derek Nexus software (Lhasa Ltd., 2014). The Cooper statistics for the DD were compared to the Published Statistics (PS) to see whether the PS could be reproduced.
- TD: the Test Dataset was specifically designed for each individual model, and consists of all chemicals from our database that were not present in the training set of that specific

model. The TD was used to evaluate the prospective predictive performance of a model for a set of unknown chemicals.

- WD: the Whole Dataset consists of all chemicals in our database, regardless whether a chemical was used in the training set of a model or not. The WD was used to compare the overall performance of the models with each other.

Our manual implementation of the Jarvis model and Enoch alerts, as well as the predictions for the model training sets obtained from the model developers directly, resulted in some differences in predictions with respect to the original published data. In these cases, the manually obtained predictions were used for further evaluations. In addition, for a few chemicals there was conflicting data on whether a chemical is considered a respiratory sensitizer or a non-sensitizer according to the individual model developers. In this case, a positive assignment in the training set of one model overruled a negative assignment for the same chemical from the other model (worst-case approach). These chemicals are commented on in supplementary table 1. The individual models were also analyzed for specific chemical structure applicability domain adequacies and/or deficiencies, using the WD.

Tiered testing approach

As a proof of concept, a tiered testing approach selecting predictions with the highest reliability was developed. Two models generating predictions on respiratory sensitization potential with the highest reliability were thus selected. The model with the highest predictive value, positive or negative, is selected for the first tier of the approach. The second tier consists of the model with the highest complementary predictive value (if tier 1 consist of the model with the highest PPV, the second tier consist of the model with the highest NPV and vice versa). Additionally, the tiered approach was optimized by compensating for some of the models deficiencies in their respective applicability domains and by complementing the tiered

approach using the adequacies of another model. Combined, the tiered approach is designed to only conclude on the respiratory sensitization potential for those substances where the models are thought to be highly reliable.

RESULTS

Reproducibility of Published Statistics

The Published Statistics (PS) from the respective publications based on the individual training sets are given for each model in table 1. The final column in Table 1 shows the actual number of predictions generated for a specific dataset over the total number of chemicals present in the specific datasets. A complete overview of the calculations leading to the Cooper statistics for each model and for each dataset can be found in supplementary tables 2 to 6.

The Development Dataset (DD) statistics for each model were compared to the respective Published Statistics (PS), except for Derek and MultiCASE for which no published predictive performance is available (Table 1). The PS of cat-SAR, Jarvis and Enoch were reproducible within 10% difference. In general, our implementation of the models or the inconsistencies of the reported predictions with those obtained directly from the model developers do not appear to influence the performance of the models significantly. The statistics from the DD are therefore used as a point of reference for further comparisons.

Evaluation of model predictive performance

The overall performance of the models was calculated using the Whole Dataset (WD) (Table 1). In all cases, the predictivity for this dataset was lower than for the DD, with the exception of the sensitivity of the Jarvis model, which remained the same for the WD compared to the DD. MultiCASE had the largest number of equivocal, out-of-domain or non-predicted

substances, 399 out of 659 (61%), because the MultiCASE implementation of the DK EPA is the only model to apply a rigorous applicability domain check. The ability of the models to predict new substances was evaluated by comparing the Test Dataset (TD) statistics to the DD (Table 1). In all cases, the TD statistics were below the DD, again with the exception of the sensitivity of Jarvis. The TD versus DD comparison is illustrated using ROC plots, showing the sensitivity versus 1-specificity (Figure 1) and for the purpose of this study more relevant PPV versus 1-NPV (Figure 2), excluding Derek as for this model only sensitivity could be calculated for the DD. In both figures, the TD performances of all SARs are closer to the diagonal than the DD. The ROC plots also show that the sensitivity of Enoch and the specificity of cat-SAR decreased most when comparing TD to PS (Figure 1). The PPV and NPV of the cat-SAR model decreased the most when comparing TD to PS (Figure 2).

Chemical applicability domain

For 26 out of the total 138 respiratory sensitizers, all models gave a positive prediction. The 26 respiratory sensitizers for which all models agreed included nine isocyanates, six anhydrides and six penicillin-like chemicals. For 76 of the 521 respiratory non-sensitizers all models agreed on a negative prediction. Of the 76 respiratory non-sensitizers on which all five models agreed, 31 contained an acetate group, 14 an alcohol group and 11 a ketone group. The other chemicals for which the models agreed on a positive or negative prediction did not share any obvious commonalities in chemical structure. There were no chemicals predicted incorrectly by all models. Therefore, it was analyzed if there were specific chemical classes predicted incorrectly by the individual models. Classes that were often incorrectly predicted by one or more models were: Acrylates, ethanolamines, quinolines, cyclohexenes and carboxylic acids. They are qualitatively discussed in the following sections. Figure 3 shows the shared chemical substructure for these chemical classes.

Acrylates

There are eight respiratory sensitizers and six non-sensitizer present in the WD containing an acrylate functionality. The acrylate and methacrylate alerts from Enoch identify all fourteen chemicals in this class as respiratory sensitizers. Therefore, this alert is defined too general. On the other hand, Derek did not predict any substances in this class to be a respiratory sensitizer, due to the absence of an acrylate alert. MultiCASE correctly classified six acrylates, but did not make a prediction for the other eight acrylates. The cat-SAR model predicted all but one acrylate to be non-sensitizers, with the one positive prediction correct. Jarvis has both correct and incorrect predictions for the acrylate respiratory sensitizers and non-sensitizers.

Ethanolamines

The WD contains seven ethanolamines that are all confirmed respiratory sensitizers. Derek is the only model that did not correctly identify these, as six out of seven ethanolamines did not contain any Derek alert. Derek does not have an ethanolamine alert and no other alerts were found for these chemicals. Derek correctly predicted the seventh ethanolamine as a respiratory sensitizer based on the 1,2-diamine alert. It can therefore be concluded that ethanolamines present a knowledge gap in the Derek model.

Quinolines

All quinolines in the WD are respiratory non-sensitizers. Both Enoch and Derek do not contain a quinoline alert and therefore all quinolines are correctly predicted to be non-sensitizers. On the other hand, the cat-SAR model incorrectly predicted all seven quinolines to be respiratory sensitizers indicating that this model is not able to make reliable predictions for

this chemical class. MultiCASE did not generate any predictions within the applicability domain for the quinolines.

Cyclohexenes

Seven out of twenty-one cyclohexenes in the WD are respiratory sensitizers. Enoch contains a cyclohexene alert and therefore predicts all twenty-one chemicals to be respiratory sensitizers. This indicates that the cyclohexene alert of Enoch is defined too broadly. Derek correctly identifies two cyclohexene respiratory sensitizers, but this is based on an anhydride alert. The other cyclohexenes are all predicted to be respiratory non-sensitizers by Derek.

Carboxylic Acids

There are 25 respiratory sensitizers and 36 non-sensitizers containing a carboxylic acid functionality in the WD. This chemical group attracted attention because all MultiCASE predictions for this chemical class were positive. A MultiCASE prediction was produced for 34 (18 respiratory sensitizers and 16 non-sensitizers) out of the total of 61 carboxylic acids. When inspecting these 34 acid-containing chemicals, only four contain a functional group that is potentially reactive towards proteins (5-Hydroxy-4-nitrobenzoic acid, Clopyralid, p-Aminobenzoic acid and anthranilic acid). The other carboxylic acid containing chemicals do not appear to have any other reactive functionality. The MultiCASE predictions for carboxylic acid containing substances are therefore considered unreliable (false positives), and apparently the applicability domain does not indicate these substances (carboxylic acids) to be out of domain.

Statistical applicability domain

When analyzing the individual Cooper statistics for each model, some models predict respiratory sensitization potential with a high reliability (indicated by a high PPV) while other

models predict the absence of respiratory sensitization potential with a high reliability (as indicated by a high NPV). Figure 4 shows that there is no single model that has both a higher PPV and NPV compared to the other models. The PPV of Derek is the highest of all models for the TD as well as the WD, which demonstrates that Derek gives the most reliable prediction of a substance being a respiratory sensitizer. MultiCASE has the highest NPV for both the WD and the TD, indicating that this model gives the most reliable prediction for respiratory non-sensitization. The PPV's and NPV's calculated to determine the best predicting models were scaled to a prior probability of 50% (an assumption that every compound in 'chemical space' has a 50% probability of being a respiratory sensitizer).

Tiered approach

Combining the chemical and statistical applicability domain information, a two-tiered approach based on the existing SAR models was developed and specifically optimized to generate only highly reliable predictions on respiratory sensitization. The model with the highest predictive value for the WD, which is the PPV of Derek, is selected to predict respiratory sensitization potential in the first tier. MultiCASE has the highest NPV for the WD and is therefore selected to predict the absence of respiratory sensitization potential in the second tier. A calculation of the PPV's and NPV's over the whole range of prior probabilities from 0 to 1 shows that Derek always has the highest PPV and MultiCASE always has the highest NPV (Supplementary figures 1 and 2). The selection of Derek and MultiCASE is therefore warranted.

The observed deficiency in the chemical applicability domain of Derek (ethanolamines) can be compensated by adding the ethanolamine alert as defined by Enoch to further improve the tiered approach. It makes sense to use the strengths of other models to complement the tier 1 model (Derek in our tiered approach). To improve tier 1 of the approach, all alerts from

Enoch that have a PPV equal to or higher than the PPV from Derek, based on at least 5 chemicals, were added. These alerts are for ethanolamines, ethylenediamines, lactams, anhydrides and diisocyanates. An additional eight chemicals are correctly predicted as respiratory sensitizers by the ethanolamine, ethylenediamine and lactam alerts. The anhydride and isocyanates alerts do not add any new predictions to tier 1 compared to Derek. No other alerts or predictions from the other SAR models are found sufficiently reliable to complement Derek.

In tier 1 of the approach, chemicals predicted positive are accepted to be respiratory sensitizers, because of the high PPV of Derek and the selected additional alerts from Enoch. These chemicals therefore do not need to proceed to tier two. Derek predicts 50 of the 659 chemicals in the WD to be respiratory sensitizers. Four of these fifty are false positive predictions (Atrazine, piperazine HCl and two drug intermediates). The added Enoch alerts to the first tier result in the correct identification of an additional eight respiratory sensitizers. In total, 54 true respiratory sensitizers and 4 false positive respiratory non-sensitizers are identified in tier 1. In tier 2, MultiCASE predicts 175 of the remaining 601 chemicals to be respiratory non-sensitizers; five of these are false negatives (Styrene, 3-carene, p-phenylene diamine, indigotine and TBTU). Exclusion of predictions for carboxylic acid compounds, which were found unreliable, was not relevant for the tiered approach. All remaining carboxylic acid compounds after tier 1 were predicted as respiratory sensitizers by MultiCASE, but only negative predictions of MultiCASE were accepted as sufficiently reliable in tier 2. The complete tiered approach is illustrated in Figure 5. 426 chemicals (65% of the WD) could not be classified with sufficient reliability by the tiered approach. Based on the 58 positive predictions in the first tier and the 175 negative predictions in the second tier, the scaled PPV and scaled NPV of the tiered approach for the WD are 98% and 92% respectively.

The prospective predictive performance of the tiered approach was also estimated using only the TD. Because the TD's from Derek and MultiCASE are not identical, the estimation of the prospective performance is based on only those chemicals that were part of both TD's. This reduced Test Dataset contains 572 chemicals, for which Derek generated 19 positive predictions. Four of those predictions are incorrect. Six additional respiratory sensitizers are identified using the additional alerts from Enoch. For the remaining 547 chemicals, MultiCASE gives 140 negative predictions of which five are false negative. Combined, the tiered approach reliably predicts 165 out of 572 (29%) chemicals from the reduced Test Dataset with a scaled PPV and NPV of 96% and 86% respectively. The tiered approach therefore performs only slightly less in predicting new chemicals compared to the performance on the WD.

Figure 6 shows the ROC plot using PPV and 1-NPV of the WD for all individual models, and the tiered approach. The orthogonal distance of the tiered approach to the diagonal is 12% larger than for MultiCASE and even 54% larger than for Derek, which were individually the best performing models for NPV and PPV. The tiered approach therefore has a significantly improved predictive performance compared to any of the individual models. This is achieved by accepting that no sufficiently reliable prediction can be obtained for 65% of the chemicals.

DISCUSSION

The results of the current study show that the predictive performances of the tested SARs are considerably lower for a new set of chemicals compared to the performance on their own training datasets. It can be concluded that no single SAR model is sufficiently reliable to predict respiratory sensitization. To examine if these models could be improved, one of the aims of this study was to identify deficiencies in the chemical and statistical applicability domains of the existing SARs. One or more models were observed to be less reliable for five

chemical classes compared to their overall performance on the WD. This knowledge can be used to improve specific models. The absence of an ethanolamine alert in Derek is an example of this. Additionally, Derek generated four false positive predictions. One is for piperazine dihydrochloride. Its chemical structure makes this substance inherently a respiratory sensitizer, as piperazine in its neutral state is considered a true respiratory sensitizer. Possibly the fact that the substance is presented as a salt reduces its (respiratory) bioavailability so much that it is not able to exert its sensitization potential in vivo. Two other false positives of Derek are precursors in the synthesis of the drug Pazopanib. They were categorized to be respiratory non-sensitizers based on negative test data in the LLNA, but predicted to be respiratory sensitizers due to the presence of a halo-diazine or triazine ring. This alert might therefore need additional evaluation, or limitation of its applicability domain. The fourth false positive of Derek is atrazine, for which no explanation could be determined. There are no apparent explanations for the five misclassifications of MultiCASE.

MultiCASE predicted several carboxylic acid containing chemicals to be respiratory sensitizers while they are non-sensitizers. One explanation for the observed misclassifications in this group is the possibility that these acid-containing chemicals act as respiratory tract irritants, instead of respiratory sensitizers, as the underlying mechanisms are difficult to distinguish in a clinical setting (Burge *et al.*, 2012). Carboxylic acid-containing chemicals are most likely a group of occupational asthmagens that do not induce hypersensitivity of the lung through an immunological reaction. Enoch *et al.* (2012) also specifically states that carboxylic acids are possibly incorrectly identified as true respiratory sensitizers instead of occupational asthmagens causing a non-immunological hypersensitivity reaction. Additionally, carboxylic acids lack a chemical mechanism through which they can covalently bind to proteins (Enoch, *et al.*, 2012), which is required for a chemical to cause sensitization (Weltzien *et al.*, 1996). The carboxylic acid containing respiratory sensitizers that were correctly predicted by

MultiCASE contained one or more other functionalities that are probably responsible for the sensitizing capabilities of those chemicals. For these reasons, theoretical predictions of respiratory sensitization potential for this class of chemicals might in general be unreliable. Additionally, more insight in the cause of occupational asthma (immunological versus non-immunological) is still needed for these substances.

The criteria used for defining respiratory non-sensitizers in the original model publications differed considerably. Jarvis and Enoch used occupational exposure limits (OELs) to define respiratory non-sensitizers while cat-SAR and MultiCASE used a mechanistic explanation that negative human skin sensitization data implied that the chemical would also be a respiratory non-sensitizer. Additionally, based on the work of Dearman et al. (2013), a large set of LLNA negative skin sensitizers were added to the dataset by us, to further increase the number of chemicals for evaluation of the SAR models. These different assignment definitions of respiratory non-sensitizers have their limitations. For example, Jarvis and Enoch implicitly assume that sensitized workers who are exposed to a LMW respiratory sensitizer will always report to a clinic. However, this might not always be true. This is demonstrated by Malo (Malo *et al.*, 2010) who showed that many workers do not apply for compensation because the financial benefits are often very low and applying for compensation may even cause a situation where workers might suffer employment prejudice. In addition, a correct diagnosis of OA is challenging, because it is difficult to distinguish a case as occupational, instead of non-occupational asthma (Aasen *et al.*, 2013).

One possible argument against the assumption that a negative result for a LMW chemical in the murine LLNA test is predictive of absence of respiratory sensitization potential is that if a chemical does not penetrate the skin, it will not elicit a response in the LLNA test. This does not guarantee that the same chemical, if inhaled, will not induce respiratory sensitization. However, a calculation of the dermal permeability rate (K_p) using MW and octanol-water

partition coefficients (Potts and Guy, 1992) for all chemicals used in this study reveals that there is no clear difference between the dermal permeability rates of LLNA negative chemicals, human skin negative chemicals, OEL negative chemicals and respiratory sensitizers (data not shown). The low correlation between sensitization potential and generic chemical properties, such as Kp, was for example also shown for aldehydes (Patlewicz *et al.*, 2001). This implies that a negative result in the LLNA is not necessarily influenced by skin absorption. Additionally, there are no examples of organic LMW respiratory sensitizers that do not also give a positive result in the LLNA. Therefore, it is a plausible assumption that chemicals in the Whole Dataset that give a negative result in the LLNA are also respiratory non-sensitizers.

A calculation of the performances of the SAR models (Supplementary tables 7 to 11) using separate sets of negative respiratory sensitization data (OEL negatives versus LLNA negatives) shows that for Enoch, Derek and cat-SAR there are no significant differences in performance between the OEL negatives and LLNA negatives. Jarvis performs better on OEL negatives, which can be explained because this model is trained using these chemicals. MultiCASE also performs better on OEL negatives, but when examining the predictions of MultiCASE, the majority of false positives within the LLNA negative dataset are carboxylic acid-containing chemicals. As discussed, theoretical predictions for carboxylic acid-containing chemicals should be taken with great care. If these predictions, for this reason, are omitted from the calculations, the performance of MultiCASE on LLNA negatives equals its performance on OEL negatives. The above mentioned considerations suggest that both sets of chemicals can be used for this study as respiratory non-sensitizers.

This paper shows that taking the most reliable predictions from two complementary models in a tiered approach yields a combined predictive performance that might be considered sufficient for regulatory purposes. However, it needs to be considered that 1) for a substantial

part of the chemicals information is too limited to reach a final conclusion on respiratory sensitization potential and 2) a decision on the mechanistic cause (immunological or non-immunological) of occupational asthma will also require more information. It is doubtful whether a higher reliability is achievable through additional testing for the 233 substances for which the SAR information was considered conclusive, using the tiered approach. However, other test methods, such as *in vitro* or *in chemico* methods, could further complement the outcomes of this tiered SAR approach and address other key events involved in respiratory sensitization. As the proposed tiered approach is not conclusive on the respiratory sensitization potential of 65% of the chemicals tested in this evaluation, additional test methods are clearly needed. Our research is currently aiming at the development of additional animal-free testing methods (*in chemico* and *in vitro*) to complement the tiered approach proposed in this paper. The use of an Integrated Testing Strategy incorporating further testing methods will hopefully allow for reliable prediction of the majority of the chemicals not predicted sufficiently reliable by the proposed SAR-based tiered approach.

SUPPLEMENTARY DATA DESCRIPTION

Two supplementary files are provided with this article; Supplementary table 1.xlsx; containing the Whole Database of 659 chemicals used to evaluate the SAR models and Supplementary datafiles.docx; containing the individual model Cooper statistics calculations, the OEL negatives versus LLNA negatives calculations and the calculation of PPV and NPV based on the whole range of prior probabilities.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Eva B. Wedeby and Nikolai G. Nikolov from the Technical University of Denmark, National Food Institute, Division of Toxicology and Risk Assessment for providing all the missing predictions from the MultiCASE model that were not available in the online database, and for critically reviewing the manuscript. The authors would also like to thank Steven Enoch for supplying the references to the individual case reports which were used to develop their structural alerts.

REFERENCES

- Aasen, T. B., Burge, P. S., Henneberger, P. K., Schlunssen, V. and Baur, X. (2013). Diagnostic approach in cases with suspected work-related asthma. *J Occup Med Toxicol* **8**, 7.
- Accelrys, *Accord for Excel v7.1* (2010). " Accord for Excel displays chemical structures and reactions, performs chemical calculations, analyzes R-groups, and queries by substructure or similarity directly within Excel." <http://accelrys.com/products/informatics/desktop-software.html>. Accessed July 2014.
- Arts, J. H. and Kuper, C. F. (2007). Animal models to test respiratory allergy of low molecular weight chemicals: a guidance. *Methods* **41**, 61-71.
- Bloemen, K., Verstraelen, S., Schoeters, G., Legiest, B. and Nemery, B. (2009). The collection and evaluation of data on incidence and severity of skin and respiratory allergy related to exposure of chemicals from non-food sources. *Report , European Commission Health & Consumer Protection Directorate*.
- Boverhof, D. R., Billington, R., Gollapudi, B. B., Hotchkiss, J. A., Krieger, S. M., Poole, A., Wiescinski, C. M. and Woolhiser, M. R. (2008). Respiratory sensitization and allergy: current research approaches and needs. *Toxicol Appl Pharmacol* **226**, 1-13.
- Burge, P. S., Moore, V. C. and Robertson, A. S. (2012). Sensitization and irritant-induced occupational asthma with latency are clinically indistinguishable. *Occup Med (Lond)* **62**, 129-133.

Cooper, J. A., 2nd, Saracci, R. and Cole, P. (1979). Describing the validity of carcinogen screening tests. *Br J Cancer* **39**, 87-89.

CSST (2013). *List of agents causing occupational asthma*. Available at:

http://www.asthme.csst.qc.ca/document/Info_Gen/AgenProf/Bernstein/BernsteinAng.htm.

Accessed July 2014.

Cunningham, A. R., Cunningham, S. L., Consoer, D. M., Moss, S. T. and Karol, M. H. (2005).

Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers. *SAR QSAR Environ Res* **16**, 273-285.

Danish (Q)SAR database. Available at: <http://qsar.food.dtu.dk/>. Accessed July 2014.

Dearman, R. J., Basketter, D. A. and Kimber, I. (2013). Inter-relationships between different classes of chemical allergens. *J Appl Toxicol* **33**, 558-565.

Dewitte, J. D., Chan-Yeung, M. and Malo, J. L. (1994). Medicolegal and compensation aspects of occupational asthma. *Eur Respir J* **7**, 969-980.

Doubilet, P. (1983). A mathematical approach to interpretation and selection of diagnostic tests. *Med Decis Making* **3**, 177-95.

Enoch, S. J., Roberts, D. W. and Cronin, M. T. (2009). Electrophilic reaction chemistry of low molecular weight respiratory sensitizers. *Chem Res Toxicol* **22**, 1447-1453.

Enoch, S. J., Roberts, D. W. and Cronin, M. T. (2010). Mechanistic Category Formation for the Prediction of Respiratory Sensitization. *Chem Res Toxicol* **23**, 1547-1555.

Enoch, S. J., Seed, M. J., Roberts, D. W., Cronin, M. T., Stocks, S. J. and Agius, R. M. (2012).

Development of mechanism-based structural alerts for respiratory sensitization hazard identification. *Chem Res Toxicol* **25**, 2490-2498.

EU (2008). *Classification, labelling and packaging of substances and mixtures*. Available at:

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:353:0001:1355:EN:PDF>. Accessed July 2014.

EU (2012). *Guidance on information requirements and chemical safety assessment Chapter R7.a:*

Endpoint specific guidance. Source: European Chemicals Agency. Available at:

http://echa.europa.eu/documents/10162/13632/information_requirements_r7a_en.pdf. Accessed July 2014.

Graham, C., Rosenkranz, H. S. and Karol, M. H. (1997). Structure-activity model of chemicals that cause human respiratory sensitization. *Regul Toxicol Pharmacol* **26**, 296-306.

Jarvis, J., Seed, M. J., Elton, R., Sawyer, L. and Agius, R. (2005). Relationship between chemical structure and the occupational asthma hazard of low molecular weight organic compounds. *Occup Environ Med* **62**, 243-250. Web-based version of the model available at:

<http://www.coeh.man.ac.uk/asthma/login.php>. Accessed July 2014.

Klopman, G. (1992). MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quantitative Structure-Activity Relationships* **11**, 176-184.

Lalko, J. F., Kimber, I., Gerberick, G. F., Foertsch, L. M., Api, A. M. and Dearman, R. J. (2012). The direct peptide reactivity assay: selectivity of chemical respiratory allergens. *Toxicol Sci* **129**, 421-431.

Lhasa Ltd. (2014). *Derek v4.0.5, Nexus v1.7.5, Knowledge Base 2014 v1.0*. "Knowledge-based software for toxicity predictions" <http://www.lhasalimited.org/products/derek-nexus.htm>. Accessed July 2014.

Malo, J. L., Lemiere, C., Gautrin, D., Labrecque, M. and Lavoie, K. (2010). Asthma and the Workplace. In *Asthma, Health and Society; A Public Health Perspective* (A. Harver and H. Kotses, Eds.), pp. 303-323. Springer, New York, NY.

Mapp, C. E., Boschetto, P., Maestrelli, P. and Fabbri, L. M. (2005). Occupational asthma. *Am J Respir Crit Care Med* **172**, 280-305.

NICEATM-ICCVAM database (2013). NICEATM LLNA Database. Available at:

<http://ntp.niehs.nih.gov/iccvam/methods/immunotox/NICEATM-LLNAdatabase-23Dec2013.xls>.

Accessed July 2014.

- Patlewicz, G., Basketter, D. A., Smith, C. K., Hotchkiss, S. A. and Roberts, D. W. (2001). Skin-sensitization structure-activity relationships for aldehydes. *Contact Dermatitis* **44**, 331-6.
- Potts, R. O. and Guy, R. H. (1992). Predicting skin permeability. *Pharm Res* **9**, 663-9.
- Pralong, J. A., Cartier, A., Vandenplas, O. and Labrecque, M. (2012). Occupational asthma: new low-molecular-weight causal agents, 2000-2010. *J Allergy (Cairo)* **2012**: Article ID 973056, 10 pages.
- Quirce, S. and Bernstein, J. A. (2011). Old and new causes of occupational asthma. *Immunol Allergy Clin North Am* **31**, 677-698.
- Quirce, S. and Sastre, J. (2011). New causes of occupational asthma. *Curr Opin Allergy Clin Immunol* **11**, 80-85.
- Rorije, E., Aldenberg, T., Buist, H., Kroese, D. and Schuurmann, G. (2013). The OSIRIS Weight of Evidence approach: ITS for skin sensitisation. *Regul Toxicol Pharmacol* **67**, 146-56.
- Sanderson, D. R. and Earnshaw, C. G. (1991). Computer prediction of possible toxic action from chemical structure: The DEREK system. *Hum Exp Toxicol*. **10**, 261-273.
- Sastre, J., Vandenplas, O. and Park, H. S. (2003). Pathogenesis of occupational asthma. *Eur Respir J* **22**, 364-373.
- Soreide, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *J Clin Pathol*. **62**, 1-5.
- Toren, K. and Blanc, P. D. (2009). Asthma caused by occupational exposures is common - a systematic analysis of estimates of the population-attributable fraction. *BMC Pulm. Med.* **9**, 7.
- UN (2007). *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*. New York and Geneva. Available at: http://www.unece.org/trans/danger/publi/ghs/ghs_rev02/02files_e.html. Accessed July 2014.
- USDHHS (1991). Publication No. 90-3042. Guidelines for Diagnosis and Treatment of Asthma. U.S. Department of Health and Human Services, National Institutes of Health.
- Vandebriel, R. J. and van Loveren, H. (2010). Non-animal sensitization testing: state-of-the-art. *Crit Rev Toxicol* **40**, 389-404.

Weltzien, H. U., Moulon, C., Martin, S., Padovan, E., Hartmann, U. and Kohler, J. (1996). T cell immune responses to haptens. Structural models for allergic and autoimmune reactions. *Toxicology* **107**, 141-151.

Table 1. Cooper statistics for all models and for all chemical datasets

Model	Dataset ^a	Sensitivity	Specificity	PPV ^b	NPV ^b	Predicted/Total
MultiCASE	PS	- ^c	- ^c	- ^c	- ^c	- ^c
	DD	1.00	1.00	1.00	1.00	74/80
	TD	0.82	0.86	0.85	0.83	192/579
	WD	0.91	0.89	0.89	0.91	266/659
cat-SAR	PS	0.94	0.87	0.88	0.94	75/80
	DD	0.97	0.97	0.97	0.97	72/80
	TD	0.80	0.32	0.54	0.61	382/579
	WD	0.85	0.39	0.58	0.72	454/659
Jarvis	PS	0.69	0.93	0.90	0.75	379/379
	DD	0.68	0.93	0.91	0.75	379/379
	TD	0.69	0.67	0.67	0.68	275/280
	WD	0.69	0.82	0.79	0.72	654/659
Enoch	PS	0.91	0.91	0.91	0.91	186/186
	DD	0.89	0.93	0.92	0.90	186/186
	TD	0.62	0.86	0.82	0.69	473/473
	WD	0.83	0.87	0.87	0.83	659/659
Derek	PS	- ^c	- ^c	- ^c	- ^c	- ^c
	DD	1.00	- ^d	- ^d	- ^d	25/25
	TD	0.19	0.99	0.96	0.55	634/634
	WD	0.33	0.99	0.98	0.60	659/659

^a Abbreviations of the datasets: PS: Published Statistics; DD: Development Dataset; TD: Test Dataset; WD: Whole Dataset.

^b The PPV and NPV are scaled to a ratio of 50:50 respiratory sensitizers : non-sensitizers.

^c There are no original published statistics available for MultiCASE and Derek.

- ^d Statistical value not calculable due to the absence of respiratory non-sensitizers for the specific dataset.

Figure 1. ROC plot using sensitivity versus 1-specificity for the Development Dataset (open markers) and the calculated statistics based on the Test Dataset (solid markers).

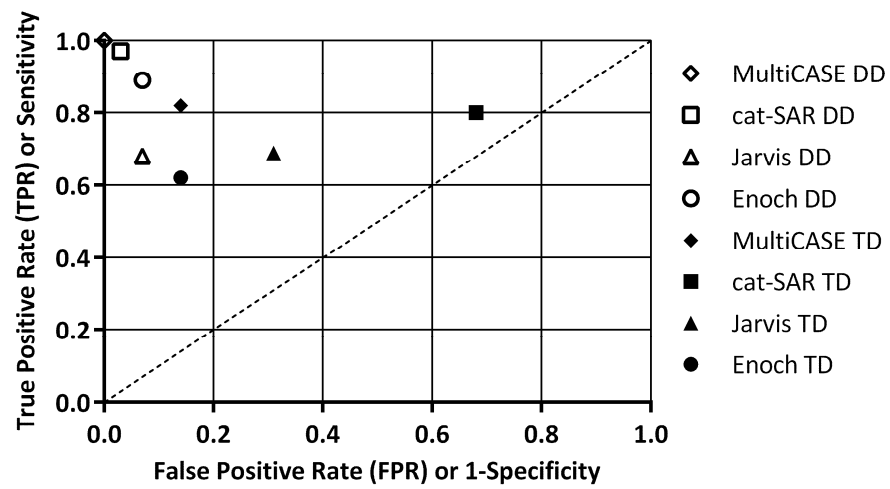
Figure 2. ROC plot using PPV versus 1-NPV for the Development Dataset (open markers) and the calculated statistics based on the Test Dataset (solid markers).

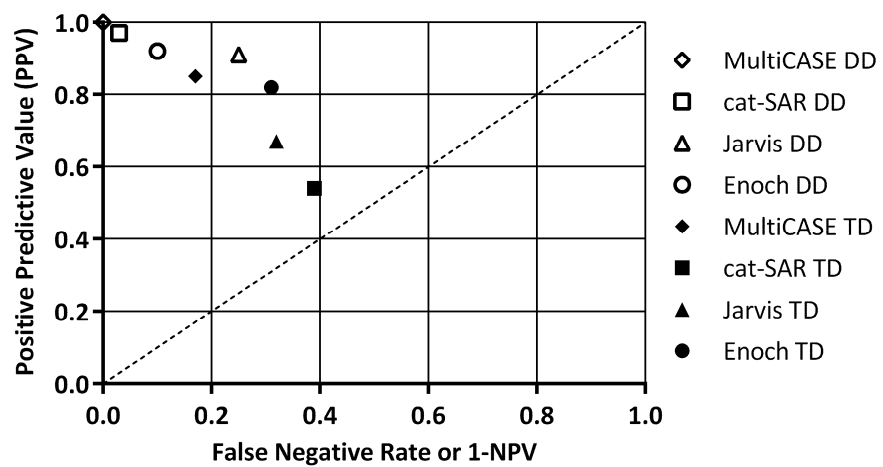
Figure 3. Chemical query structures of the classes observed to be predicted with low reliability by one or more of the SARs. R = C or H.

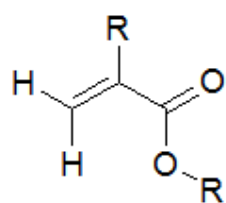
Figure 4. The calculated PPV and NPV based on the Whole Dataset for each model.

Figure 5. Illustration of a tiered approach using the model with the highest PPV and the model with the highest NPV. The identified deficiencies and adequacies in the applicability domain of the selected models are also taken into account.

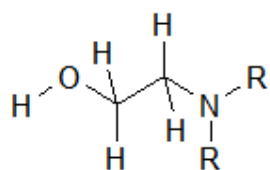
Figure 6. Performance of the individual models and the tiered approach based on the Whole Dataset.



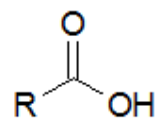




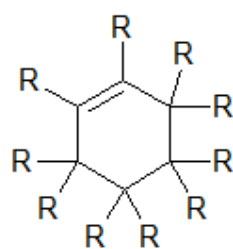
Acrylates



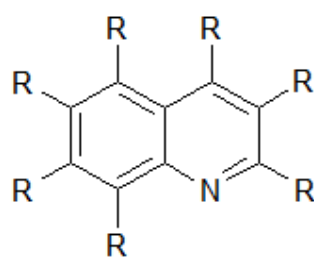
Ethanolamines



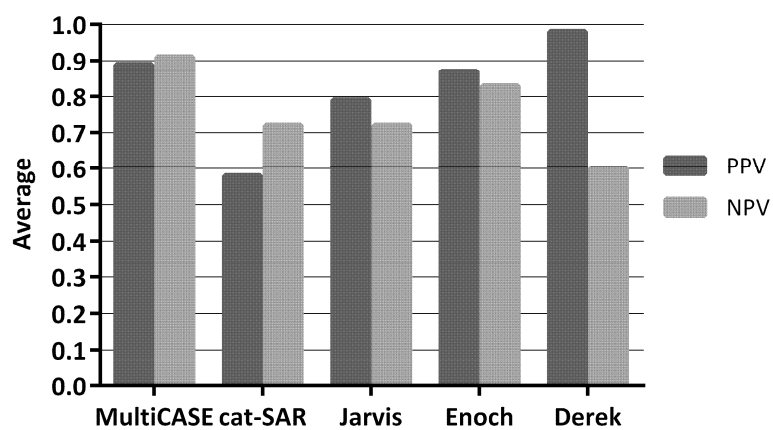
Carboxylic acids

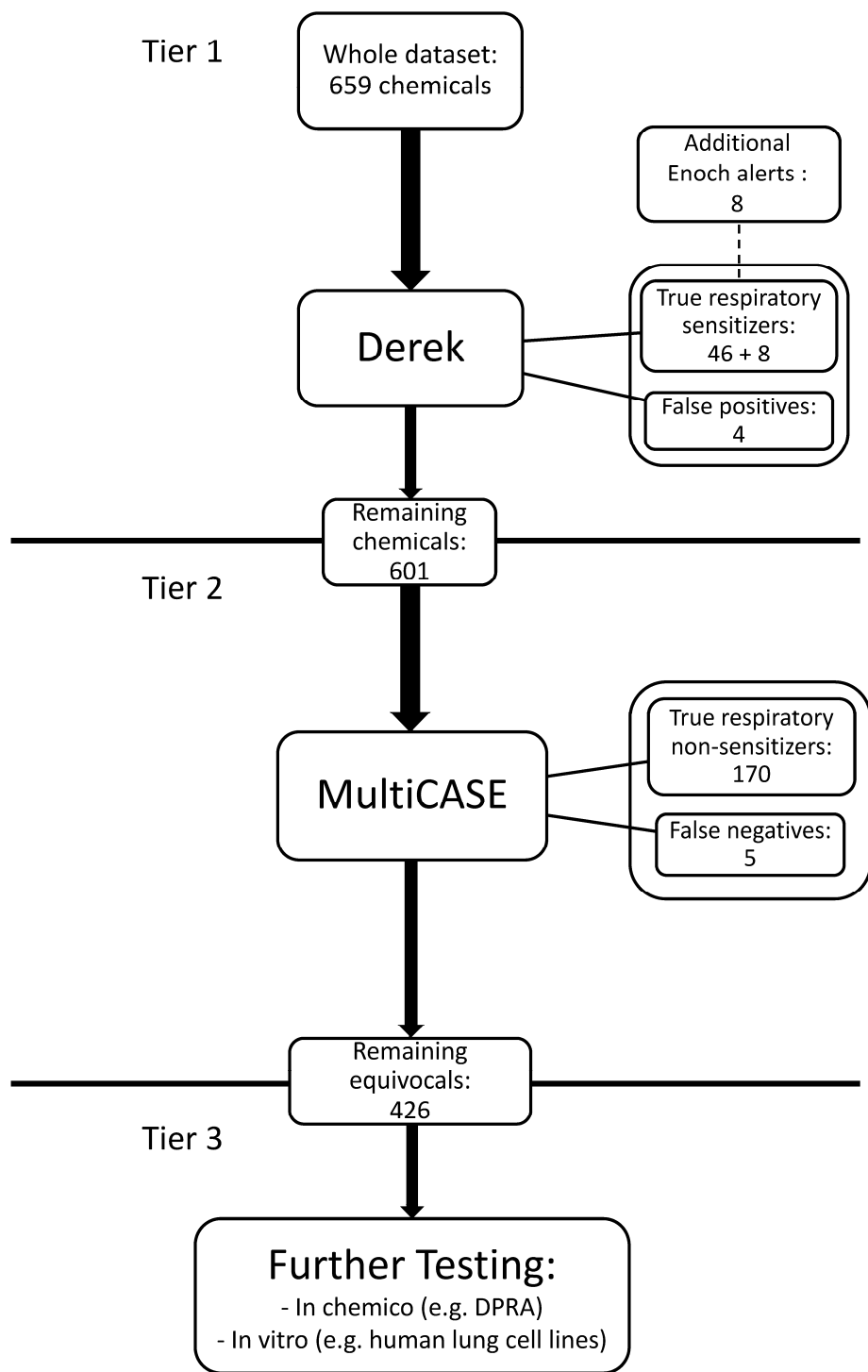


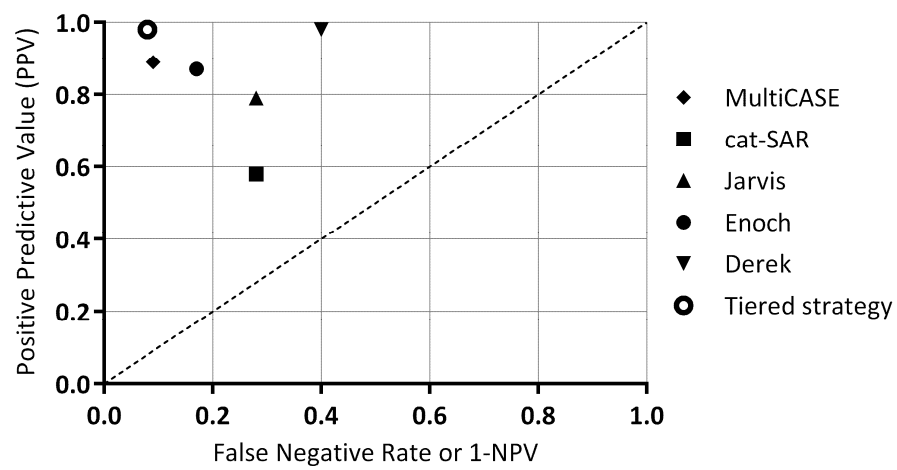
Cyclohexenes



Quinolines







Supplemental Data

For

Evaluation of in silico models for the identification of respiratory sensitizers

Sander Dik, Janine Ezendam, Albert R. Cunningham, C. Alex Carrasquer, Henk van Loveren,
and Emiel Rorije^{*}

Supplementary table 1; Whole Dataset

This table shows all 659 chemicals in the Whole Dataset and provides all information used for this study. Because of its size, it is separately attached to the paper as a supplementary file.

Detailed overview of all predictions of all models and for each dataset

The following five tables (Supplementary tables 2-6) contain the truth tables for all models with all datasets. The Cooper statistics calculated from these truth tables are given in Table 1 of the main article.

Supplementary table 2; Truth table for MultiCASE

		Respiratory sensitizer?		Total number of predictions
		+	-	
PS ^a	MultiCASE +	-	-	-
	MultiCASE -	-	-	
DD	MultiCASE +	36	0	74
	MultiCASE -	0	38	
TD	MultiCASE +	27	21	192
	MultiCASE -	6	138	
WD	MultiCASE +	63	21	266 ^b
	MultiCASE -	6	176	

^a There are no original published statistics available for MultiCASE.

^b MultiCASE applies an applicability domain, hence the total number of predictions for the WD is not 659.

Supplementary table 3; Truth table for cat-SAR

		Respiratory sensitizer?		Total number of predictions
		+	-	
PS	cat-SAR +	34	5	75
	cat-SAR -	2	34	
DD	cat-SAR +	35	1	72
	cat-SAR -	1	35	
TD	cat-SAR +	63	207	382
	cat-SAR -	16	96	
WD	cat-SAR +	98	208	454 ^a
	cat-SAR -	17	131	

^a cat-SAR does not make default predictions based on lack of evidence, so not all chemicals from the WD are predicted by cat-SAR. Therefore, the total number of predictions for the WD is not 659.

Supplementary table 4; Truth table for Jarvis

		Respiratory sensitizer?		Total number of predictions
		+	-	
PS	Jarvis +	54	22	379
	Jarvis -	24	279	
DD	Jarvis +	56	21	379
	Jarvis -	26	276	
TD	Jarvis +	38	73	275
	Jarvis -	17	147	
WD	Jarvis +	94	94	654 ^a
	Jarvis -	43	423	

^a The Jarvis model was not able to generate a prediction for 5 chemicals, so the total number of predictions for the WD is 654, not 659.

Supplementary table 5; Truth table for Enoch

		Respiratory sensitizer?		Total number of predictions
		+	-	
PS	Enoch +	95	7	186
	Enoch -	9	75	
DD	Enoch +	93	6	186
	Enoch -	11	76	
TD	Enoch +	21	60	473
	Enoch -	13	379	
WD	Enoch +	114	66	659
	Enoch -	24	455	

Supplementary table 6; Truth table for Derek

		Respiratory sensitizer?		Total number of predictions
		+	-	
PS ^a	Derek +	*	*	*
	Derek -	*	*	
DD	Derek +	25	0	25
	Derek -	0	0	
TD	Derek +	21	4	634
	Derek -	92	517	
WD	Derek +	46	4	659
	Derek -	92	517	

^a There are no original published statistics available for Derek.

Discussion on using OEL negatives versus LLNA negatives

The statistical differences in sensitivity, specificity, PPV and NPV for all models based on LLNA negatives and/or OEL negatives is given in the following five tables (supplementary tables 7-11). They each show three different datasets:

- 1) Whole Dataset, the same statistics as can be found in the main article (Table 1).
- 2) Reduced dataset based on LLNA negatives that are not also OEL negatives. In other words, this is a dataset containing all respiratory sensitizers, human skin non-sensitizers and LLNA negative non-sensitizers.
- 3) Reduced dataset based on OEL negatives that are not also LLNA negatives. In other words, this is a dataset containing all respiratory sensitizers, human skin non-sensitizers and OEL negative non-sensitizers.

Supplementary table 7; Comparison of LLNA versus OEL Cooper statistics for MultiCASE

	MultiCASE Whole Dataset	MultiCASE OEL negatives	MultiCASE LLNA negatives
Sensitivity	0.91	0.91	0.91
Specificity	0.89	0.95	0.84
PPV	0.89	0.95	0.85
NPV	0.91	0.92	0.90

The sensitivity and NPV do not differ between the datasets. The specificity and PPV are 10% lower for the reduced dataset based on LLNA negatives compared to the reduced dataset based on OEL negatives.

Supplementary table 8; Comparison of LLNA versus OEL Cooper statistics for cat-SAR

	cat-SAR Whole Dataset	cat-SAR OEL negatives	cat-SAR LLNA negatives
Sensitivity	0.85	0.85	0.85
Specificity	0.39	0.49	0.39
PPV	0.58	0.62	0.58
NPV	0.72	0.77	0.72

The sensitivity, PPV and NPV do not differ significantly between the datasets. The specificity is 10% lower for the reduced dataset based on LLNA negatives compared to the reduced dataset based on OEL negatives.

Supplementary table 9; Comparison of LLNA versus OEL Cooper statistics for Jarvis

	Jarvis Whole Dataset	Jarvis OEL negatives	Jarvis LLNA negatives
Sensitivity	0.69	0.69	0.68
Specificity	0.82	0.90	0.72
PPV	0.79	0.87	0.71
NPV	0.72	0.74	0.69

The sensitivity and NPV do not differ significantly between the datasets. The specificity and PPV are considerably lower for the reduced dataset based on LLNA negatives compared to the reduced dataset based on OEL negatives.

Supplementary table 10; Comparison of LLNA versus OEL Cooper statistics for Enoch

	Enoch Whole Dataset	Enoch OEL negatives	Enoch LLNA negatives
Sensitivity	0.83	0.83	0.83
Specificity	0.87	0.88	0.87
PPV	0.87	0.87	0.87
NPV	0.83	0.84	0.84

There are no differences between the sensitivity, specificity, PPV and NPV for any of the datasets.

Supplementary table 11; Comparison of LLNA versus OEL Cooper statistics for Derek

	Derek Whole Dataset	Derek OEL negatives	Derek LLNA negatives
Sensitivity	0.33	0.33	0.33
Specificity	0.99	1.00	0.99
PPV	0.98	0.99	0.96
NPV	0.60	0.60	0.60

There are no differences between the sensitivity, specificity, PPV and NPV for any of the datasets.

Discussion on using a fixed prior probability of 50% to calculate PPV and NPV

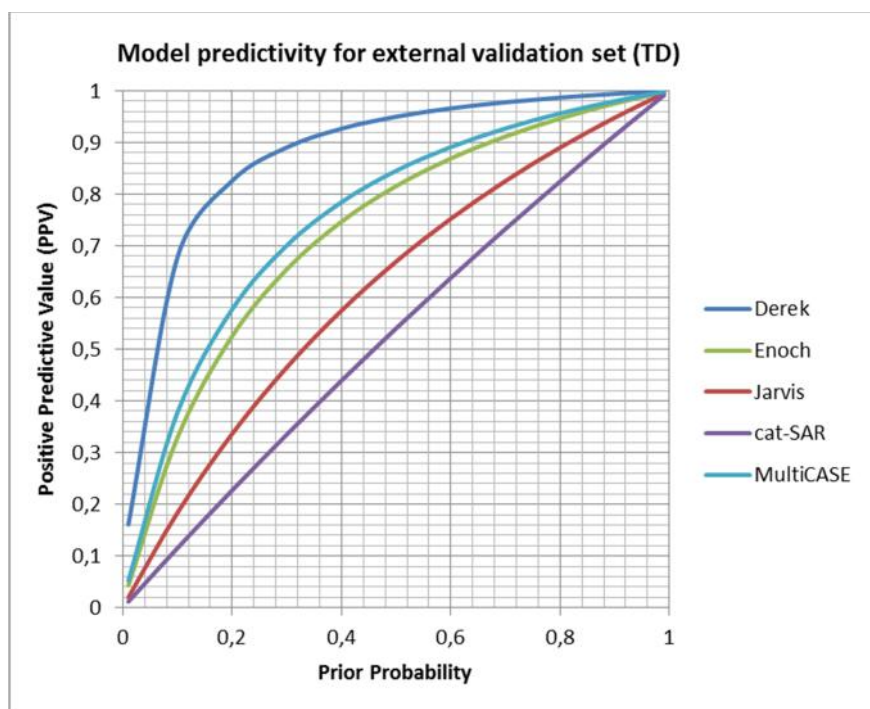
The following two figures show the PPV (Supplementary figure 1) and NPV (Supplementary figure 2) of all models calculated for all prior probabilities ranging from 0 to 1. The PPV and NPV were calculated as a function of the prior probability, sensitivity and specificity of a model^a (See below). The sensitivity and specificity are based on the Test Dataset of a model, which can be found in Table 1 of the main paper.

$$\text{PPV} = (\text{Sensitivity} * \text{Prior probability}) / ((\text{Sensitivity} * \text{Prior probability}) + (1 - \text{Specificity}) * (1 - \text{Prior probability}))$$

$$\text{NPV} = (\text{Specificity} * (1 - \text{Prior probability})) / (\text{Specificity} * (1 - \text{Prior probability}) + (1 - \text{Sensitivity}) * \text{Prior probability})$$

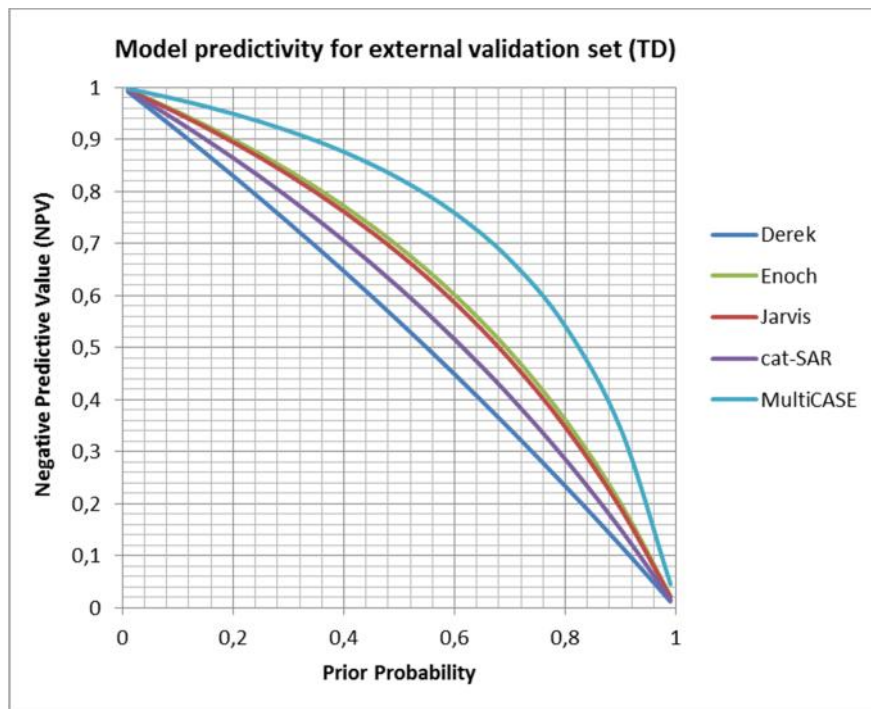
^a Rorije, E., Aldenberg, T., Buist, H., Kroese, D. and Schuurmann, G. (2013). The OSIRIS Weight of Evidence approach: ITS for skin sensitisation. *Regul Toxicol Pharmacol* **67**, 146-56.

Supplementary figure 1; Graph of PPV for each model using a prior probability of 0 to 1



As can be seen from the graph, Derek has the highest PPV of all models, regardless of the prior probability. Therefore, the selection of Derek in tier 1 of the tiered approach is warranted, as the selection of the model does not depend on the chosen prior probability of 50%.

Supplementary figure 2; Graph of PPV for each model using a prior probability of 0 to 1



As can be seen from the graph, MultiCASE has the highest NPV of all models, regardless of the prior probability. Therefore, the selection of MultiCASE in tier 2 of the tiered approach is warranted, as the selection of the model does not depend on the chosen prior probability of 50%.