Occupational Medicine 2010;**60:**115–120
Advance Access publication on 2 December 2009 doi:10.1093/occmed/kqp168

Further validation of computer-based prediction of chemical asthma hazard

Martin Seed and Raymond Agius

Occupational and Environmental Health Research Group, University of Manchester, Room C4.13, Ellen Wilkinson Building, Oxford Road, Manchester M13 9PL, UK.

Correspondence to: Martin Seed, Occupational and Environmental Health Research Group, University of Manchester, Room C4.20, Ellen Wilkinson Building, Oxford Road, Manchester M13 9PL, UK. Tel: +44 (0)161 275 5524; fax: +44 (0)161 275 5595; e-mail: martin.seed@manchester.ac.uk

Background There is no agreed protocol for the prediction of low molecular weight (LMW) respiratory sensitizers. This creates challenges for occupational physicians responsible for the health of workforces using novel chemicals and respiratory physicians investigating cases of occupational asthma caused by novel asthmagens.

Aims To iterate the external validation of a previously published quantitative structure–activity relationship (QSAR) model for the prediction of novel chemical respiratory sensitizers and to better characterize its predictive accuracy.

Methods An external validation set of control chemicals was identified from the Australian Hazardous Substances Information System. An external validation set of asthmagenic chemicals was identified by a thorough search of the peer-reviewed literature from January 1995 onwards using the Medline database. The QSAR model was used to determine an 'asthma hazard index' (between 0 and 1) for each chemical.

Results A total of 28 external validation asthmagens and 129 control chemicals were identified. The area under the receiver operating characteristic (ROC) curve for the model's ability to distinguish asthmagens from controls was 0.87 (95% CI 0.76–0.97). Using a cut-off hazard index of 0.5 resulted in sensitivity of 79% and specificity of 93%. For prior probability ranging from 1:300 to 1:100, the negative predictive value (NPV) was 1 and positive predictive value (PPV) 0.04–0.1 while for prior probability ranging from 1:20 to 1:3, the NPV was 0.91–0.99 and PPV 0.39–0.85.

Conclusions The ROC curve for this QSAR demonstrates good global predictive power for distinguishing asthmagenic from non-asthmagenic LMW organic compounds. Potential for utilization by occupational and respiratory physicians is evident from its predictive values.

Key words Novel asthmagen; occupational asthma; QSAR; respiratory sensitizer.

Introduction

Occupational asthma is the commonest occupational respiratory disease arising from recent exposures [1,2]. New causal agents are constantly being recognized [3,4] and therefore this can present a challenge for preventative measures as well as for causal attribution in a diagnostic context. Regulatory toxicology and official schedules of recognized causal agents are of limited, if any, value when trying to prevent or diagnose asthma from specific novel causes. However, having a chemical structure related to substances known to cause respiratory hypersensitivity is one item of evidence for respiratory sensitization potential in humans [5]. A preliminary exploration of this concept for low molecular weight (LMW) organic chemicals revealed that certain chemical substructures, such as unsaturated bonds involving carbon and nitrogen heteroatoms, were over-represented in respiratory sensitizers as compared to control chemicals [6]. Another statistical finding of this initial pilot study was the greater mean number of reactive or functional groups present in asthmagenic compared to non-asthmagenic compounds.

These observations were used to develop a quantitative structure–activity relationship (QSAR) model for the prediction of asthma hazard of LMW organic compounds [7]. This model estimates the probability that an unknown chemical has asthmagenic potential, termed

'asthma hazard index', from a logistic regression analysis of a range of molecular descriptors. The model's initial external validation showed promise in distinguishing asthmagenic from non-asthmagenic compounds with a sensitivity of 86% and specificity of 99%. As further novel causes of occupational asthma are continually reported in the literature, it is appropriate to update periodically the external validation data set as part of an iterative process of validation. These novel asthmagens can also be considered for inclusion in expanded learning data sets for future development of refined versions of this QSAR model.

In practice, the important indicators of the usefulness of a predictive model are its positive predictive value (PPV) and negative predictive value (NPV). These vary with the prior probability that a chemical is asthmagenic, which depends on the context of use. Thus, in the scenario where an occupational physician is assessing the risk of occupational asthma occurring among a workforce exposed to a novel chemical, there may be no prior data for this hazard, i.e. the prior probability that the chemical is asthmagenic is the probability that any chemical selected at random from the thousands of known chemical structures is asthmagenic. However, when a human case of occupational asthma has occurred and a respiratory physician is considering exposure to a novel chemical as the cause, the prior probability is likely to be higher and the predictive values different.

Therefore, in this iterative validation, we demonstrate the global predictive accuracy of the asthma hazard QSAR model by plotting its receiver operating characteristic (ROC) curve. The predictive values are estimated for the use of the model by an occupational physician in risk management as well as by the specialist occupational or respiratory physician trying to identify the specific causative agent in a case of occupational asthma when the patient has been exposed to several chemicals that have not previously been recognized as asthmagenic.

Methods

An external validation set of control chemicals was identified from the Australian Hazardous Substances Information System (HSIS). Since the previous validation [7] used controls from the UK (HSE 2002) and US (ACGIH 2002) tables of occupational exposure limits, the current equivalent Australian tables [8] offered a source of fresh non-asthmagenic chemicals for use in an expanded control set. A chemical was selected from the HSIS Consolidated List (CAS Number Index) for inclusion in the validation control data set if:

(i) it could be identified by CAS Number and specific chemical structure
(ii) it had been assigned a time-weighted average exposure limit
(iii) it had not featured in either the model development control or asthmagen data sets
(iv) it was an organic chemical, i.e. containing at least one carbon atom, with a molecular weight below 1000 Da.

An external validation set of asthmagenic chemicals was identified by a thorough search of the peer-reviewed literature between January 1995 and December 2008 using the Medline database. In order to identify as many published case reports of occupational asthma as possible, both keyword- and textword-based searches were performed. Examples of search terms were 'respiratory sensitization', 'occupational asthma', 'asthma' and 'chemical'. A chemical was included in the asthmagen validation data set if:

(i) it could be identified by CAS Number and specific chemical structure
(ii) there had been at least one peer-reviewed case report of physician-diagnosed occupational asthma attributed to that chemical following a latent period of exposure
(iii) it had not featured in the model development asthmagen data set
(v) the chemical was organic with molecular weight below 1000 Da.

For each chemical in both the control and asthmagen validation data sets, the molecular structure was identified and obtained electronically in .molfile format [9]. The .molfile representation of each structure was then entered into the asthma hazard QSAR which is freely available on the Internet through URL: http://www.medicine.manchester.ac.uk/oeh/research/workrelatedillhealth/asthma/.

For each chemical an asthma hazard index between 0 and 1 was thereby obtained. A hazard index cut-off point of 0.5, as previously described [7], was used.

Each chemical was then tabulated with its corresponding asthma hazard index. A ROC plot of sensitivity against 1-specificity was produced using each of these determined hazard indices as the cut-off point [10]. This was achieved using the Stata software program version 9.

The model's sensitivity and specificity were then determined from these data and its PPV and NPV calculated over a range of prior probabilities of a chemical being asthmagenic. This range was selected to represent two very different contexts in which the model is of potential value. For toxicological screening of random chemicals for asthma hazard, the prior probability was estimated to range from 1 in 300 to 1 in 100. This approximate lower estimate of prior probability was determined by considering that of >30 000 chemicals for which toxicity data are required under **R**egistration, **E**valuation, **A**uthorisation & restriction of **Ch**emicals (REACH)

legislation, only ~100 have been reported to cause one or more cases of human asthma. However, since it is likely that a considerably greater proportion of chemicals are truly asthmagenic, 1 in 100 (rather than 1 in 300) was used as an upper estimate of the prior probability that a chemical selected at random is asthmagenic.

The other context of use for which predictive values were estimated is the situation where a physician is considering a number of chemicals as the possible novel causative agent in a case of occupational asthma. A typical number of workplace chemicals to be considered as possible causal agents in an individual patient might range from 3 to 20; hence, predictive values are determined for the prior probability of a particular chemical being the causative agent ranging from 1 in 20 to 1 in 3.

The formulae used to determine the predictive values for this hazard prediction model were identical to those that are well recognized for clinical diagnostic tests [11] with the prior probability of a chemical being asthmagenic being equivalent to 'disease prevalence'.

For this study, no ethical approval was needed as no clinical data were used that were not already in the public domain.

Results

One hundred and twenty-nine validation control chemicals that met the criteria listed above, with a mean asthma hazard index of 0.13, were identified from the HSIS. One hundred and twenty of these 129 chemicals had an asthma hazard index <0.5, which has previously been used as a cut-off point for distinguishing asthmagens from non-asthmagens. Thus, the estimated specificity of the QSAR model is 93% for this cut-off point. The nine 'false positives' were amitrole, pindone, disulfiram, *n*-ethylmorpholine, 2-*n*-dibutylaminoethanol, calcium cyanamide, propranolol, atrazine and metribuzin. Thirty-two of these 129 control chemicals (25%) had featured in the initial published validation [7]. Of the 97 controls that were not included in the initial validation, 89 had a hazard index <0.5, thus the specificity was also 93% for the novel control set.

A further seven novel asthmagenic chemicals were identified from the post-1995 peer-reviewed literature that had not appeared in the initial validation set [7]. These compounds were thiamine, eugenol, ortho-phthalaldehyde, sevoflurane, thiamphenicol, 2,4-dichloro-5-chlorosulphonyl-benzoic acid and 3-amino-5-mercapto-1,2,4-triazole. The final updated validation set comprised 28 asthmagenic chemicals as listed in Table 1, whose mean asthma hazard index was 0.77. Twenty-two of these 28 chemicals had an asthma hazard index greater than the arbitrary cut-off point of 0.5. Thus, the estimated sensitivity of the QSAR model is 79% for this cut-off point.

The ROC curve for these data is shown in Figure 1. The area under the curve is 0.87 (95% CI 0.76–0.97).

Table 1. Asthma hazard QSAR external validation asthmagenic chemicals

CAS No.	Chemical name	Hazard index
52-26-6	Morphine hydrochloride	**0.08**
59-43-8	Thiamine	0.95
97-53-0	Eugenol	**0.01**
111-42-2	Diethanolamine	1
115-27-5	Chlorendic anhydride	0.86
485-47-2	Ninhydrin	1
643-79-8	Ortho-phthalaldehyde	0.73
860-22-0	Indigotine	0.92
2451-62-9	Triglycidyl isocyanurate	1
2634-33-5	1,2-Benzisothiazolin-3-one	**0.14**
3740-18-9	2,4-Dichloro-5-chlorosulphonyl-benzoic acid	**0.33**
7696-12-0	Tetramethrin	0.97
8001-54-5	Benzalkonium chloride	0.98
15318-45-3	Thiamphenicol	0.99
16691-43-3	3-Amino-5-mercapto-1,2,4-triazole	0.71
19438-64-3	Methyltetrahydrophthalic anhydride	0.95
24447-78-7	Ethoxylated bisphenol A diacrylate	0.95
28523-86-6	Sevoflurane	**0**
38661-72-2	1,3-Bis(isocyanatomethyl)cyclohexane	1
52185-43-0	Piperidinyl chlorotriazine derivative	1
59703-84-3	Piperacillin	1
64265-57-2	Polyfunctional aziridine	1
65271-80-9	Mitoxanthrone	0.95
66592-87-8	Cefadroxil	1
72558-82-8	Ceftazidime	1
79622-59-6	Fluazinam	**0**
82547-81-7	Cefteram pivoxil	1
25700-67-6	TBTU	0.96

Hazard indices for compounds misclassified as non-asthmagenic by the QSAR model are in bold.

Figure 2 illustrates the variation in PPV and NPV calculated using sensitivity and specificity determined for a cut-off hazard index of 0.5 at a range of prior probabilities that a given chemical is asthmagenic.

Discussion

This further external validation of the QSAR model of Jarvis *et al.* [7] incorporates the first ever determination of a ROC curve for this model (or indeed of any human respiratory sensitization QSAR). In summary, it suggests that the model can perform very well in a negative predictive manner as part of a strategy to prevent asthma. The model can also perform very well in a situation of high prior probability of asthma, e.g. where a clinical decision has to be made as to what suspect agent should be used in a bronchial challenge test.

ROC curves are often used in the statistical evaluation of diagnostic tests in clinical medicine [10]. The area under the ROC curve of 0.87 indicates a good global

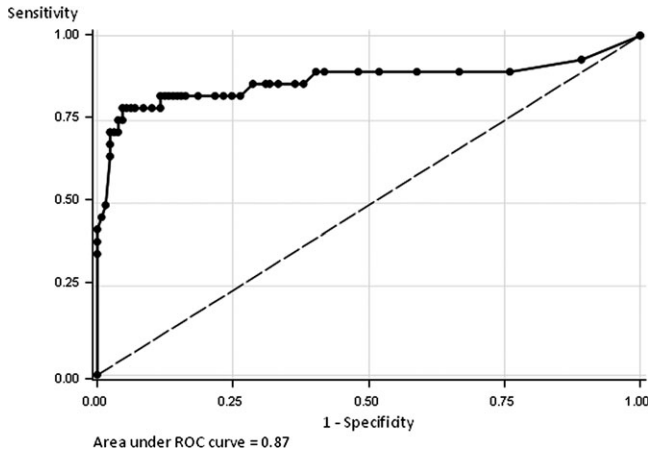

Figure 1. ROC curve for the second external validation of asthma hazard QSAR developed by Jarvis *et al.* [7].

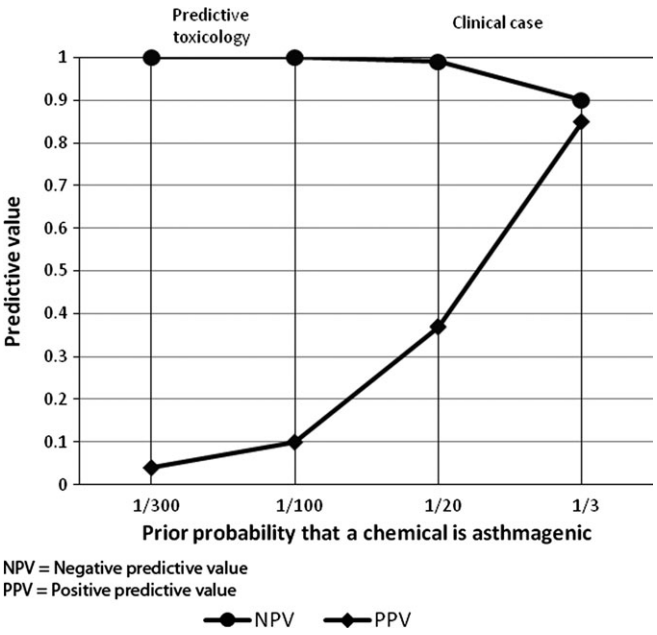

Figure 2. Variation in QSAR predictive values with context of use.

predictive power for this asthma hazard QSAR model [12]. When a hazard index of 0.5 is used as the cut-off point, the estimated specificity is 93%. This specificity may be an underestimate as it is possible that some of the controls with a hazard index >0.5 have asthmagenic potential but the required human exposure circumstances have not occurred or cases of asthma have not yet been reported. Atrazine, for example, could be an asthmagen as it is a herbicide associated with an increased odds ratio of wheeze in farmers as well as a dose–response relationship [13]. Amitrole, whose alternative name is 3-amino-1,2,4-triazole, is not without pulmonary toxicity as a case of severe alveolitis has been attributed to inhalation of this herbicide [14]. Moreover, an outbreak of occupational asthma has been attributed to a chemical with closely related structure, 3-amino-5-mercapto-1,2,4-triazole [15].

The estimated sensitivity of this model using the same cut-off hazard index of 0.5 is 79%. Only seven further case reports of occupational asthma attributed to novel LMW organic asthmagenic compounds could be identified from the literature between January 2004 and December 2008. Three of these additional seven chemicals were false negatives: eugenol, 2,4-dichloro-5-chlorosulphonyl-benzoic acid and sevoflurane. The latter two compounds both contain multiple halogen atoms, which according to the current model make them less likely to be asthmagenic. If these seven asthmagens were taken alone to validate this model, then the estimated sensitivity would only be 57%. However, small subsets are more susceptible to random error and there is no reason to suspect that these most recently published asthmagens are any more representative than the 21 asthmagens used for the initial validation. Thus, the cumulative sensitivity of 79% is a more statistically reliable estimate of the model's sensitivity.

One important potential application of this model is as a tool to assist hazard identification in the initial stages of risk assessment for occupational asthma. The validity of a predictive model for toxicological screening is determined by its predictive values. As it is not possible to know the absolute proportion of all chemicals used in industry that truly have asthmagenic potential, we have estimated this to be very approximately to the order 1 in 300. Figure 2 illustrates that the QSAR model's NPV in this context is one. The PPV is considerably lower and therefore further *in vitro* or *in vivo* data might be necessary to state with confidence that a chemical has asthmagenic potential if it has an asthma hazard index >0.5. However, for the majority of novel chemicals which have an asthma hazard index <0.5, an occupational physician could reassure workers and risk managers with some degree of confidence that asthma is a very unlikely health outcome of exposure. Similar reasoning has led to suggested utilization of a computer-based approach as the initial step in an efficient regulatory screening protocol for thousands of chemicals requiring assessment for respiratory sensitization hazard [16].

For a respiratory or occupational physician investigating the cause of a case of occupational asthma, an asthma hazard index >0.5 has more value than for an occupational physician advising on risk management in the workplace. The prior probability that a chemical being considered as the cause of an asthma case is truly asthmagenic would vary depending on the number of chemicals being considered from the exposure history. Thus, when the cause has been narrowed down to 20 possible compounds, the PPV is 0.39 increasing to 0.85 when only three chemicals are under consideration. Thus, a 'positive' (>0.5) asthma hazard index may help to corroborate causal attribution to a specific chemical entity

or identify the most likely candidate chemicals on which to perform definitive clinical investigations such as bronchial challenge testing. This is illustrated in a case report published since this period of external validation in which the authors used the QSAR-generated asthma hazard index of 0.94 for dodecanedioic acid as corroborating evidence that they have described a novel asthmagen [17]. This tool is not proposed as a substitute for full clinical assessment (including bronchial provocation tests if necessary) in cases of suspected novel causes of asthma.

The predictive accuracy of a QSAR is a function of its learning data sets. As the range of confirmed asthmagenic chemicals expands with the continuing recognition of novel causes, the learning data sets can be expanded to generate refined models which could be expected to have improved predictive values. It may also be important in future versions of this model to reconsider inclusion criteria for both learning data set asthmagens and controls. For example, it may be argued that some of the asthmagens used in the model are not truly asthmagenic, particularly those based on a solitary case report without a positive bronchial challenge test as supportive evidence. A more robust model could be generated by including only those chemicals whose asthmagenicity has been confirmed by a positive bronchial challenge test in at least two different centres.

In future work we will explore further sophistications in learning algorithms and in representations of molecular structure. It is possible that automated construction of these structure–activity relationships in a workflow environment [18] will provide the most efficient mechanism to explore combinations of descriptors and learning algorithms. We also aim to consider the influence of structural similarity between molecules in the set. Such 'redundancy' in the data can focus learning on over-represented structural themes, such that prediction is easiest for the most common molecular types. This problem has been little considered in structure–activity relationship modelling to date [19] but can be avoided rather simply by picking representative molecules from groups of similar molecules defined by clustering. With these developments, the information in this important set of asthmagenic and control compounds can be utilized to create a better prediction model and to further develop mechanistic hypotheses.

The ROC curve from this iterative external validation of the QSAR model developed by Jarvis *et al.* [7] demonstrates good global predictive power for distinguishing asthmagenic from non-asthmagenic LMW organic compounds. We have illustrated how the variation in its predictive values can be exploited to suit the purposes of both occupational and respiratory physicians. An evaluation of how the current version of this QSAR can help respiratory physicians identify novel causes of occupational asthma is planned. Plans to refine the model in order to improve its predictive power have also been outlined.

Key points

- A computer-based quantitative structure–activity relationship model has good global discriminatory power for distinguishing asthmagenic from non-asthmagenic low molecular weight organic chemicals.
- For a novel chemical introduced into the workplace, the high negative predictive value of a quantitative structure–activity relationship model allows an occupational physician to be confident that no specific respiratory precaution is likely to be required for the majority of chemicals that are not respiratory sensitizers.
- For a chemical that is being considered in the clinic as a possible novel asthmagen, the positive predictive value can be utilized by the investigating physician.

Acknowledgements

The authors would like to acknowledge the advice and comments of Dr Paul Dobson and Dr Yogendra Patel of the Manchester Interdisciplinary Biocentre, Department of Chemistry, University of Manchester, in the preparation of the manuscript.

Conflicts of interest

None declared.

References

1. Health and Safety Executive. *Statistics. Table THORR04. Work-related and occupational respiratory disease: estimated number of cases by country (a) and diagnostic category, 2006–2008.* http://www.hse.gov.uk/statistics/tables/thorr04.htm (14 November 2009, date last accessed).
2. Kogevinas M, Zock JP, Jarvis D *et al.* Exposure to substances in the workplace and new-onset asthma: an international prospective population-based study (ECRHS-II). *Lancet* 2007;**370:**336–341.
3. Drought VJ, Francis HC, Niven RM, Burge PS. Occupational asthma induced by thiamine in a vitamin supplement for breakfast cereals. *Allergy* 2005;**60:**1213–1214.
4. Ye YM, Kim HM, Suh CH, Nahm DH, Park HS. Three cases of occupational asthma induced by thiamphenicol: detection of serum-specific IgE. *Allergy* 2006;**61:**394–395.
5. United Nations. *Globally Harmonised System of Classification and Labelling of Chemicals (GHS)*. 2nd revised edn, 2007. http://www.unece.org/trans/danger/publi/ghs/ghs_rev02/02files_e.html (14 November 2009, date last accessed).
6. Agius RM, Elton RA, Sawyer L, Taylor P. Occupational asthma and the chemical properties of low molecular weight organic substances. *Occup Med (Lond)* 1994;**44:**34–36.

7. Jarvis J, Seed MJ, Elton R, Sawyer L, Agius R. Relationship between chemical structure and the occupational asthma hazard of low molecular weight organic compounds. *Occup Environ Med* 2005;**62:**243–250.
8. Australian Government. Department of Education, Employment and Workplace Relations. Office of the Australian Safety and Compensation Council. *Hazardous Substances Information System. Consolidated Lists. CAS Number Index*. http://hsis.ascc.gov.au/TheList.aspx. (Chemicals identified 10 February 2009) (23 June 2009, date last accessed.).
9. Dalby A, Nourse JG, Hounshell WD *et al*. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Comput Sci* 1992;**32:**244–255.
10. Altman DG, Bland JM. Statistics notes: diagnostic tests 3: receiver operating characteristic plots. *Br Med J* 1994;**309:**188.
11. Altman DG, Bland JM. Statistics notes: diagnostic tests 2: predictive values. *Br Med J* 1994;**309:**102.
12. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J Chem Inf Model* 2006;**4:**193–200.
13. Hoppin JA, Umbach DM, London SJ, Alavanja MCR, Sandler DP. Chemical predictors of wheeze among farmer pesticide applicators in the agricultural health study. *Am J Crit Care Med* 2002;**165:**683–689.
14. Balkisson R, Murray D, Hoffstein V. Alveolar damage due to inhalation of amitrole-containing herbicide. *Chest* 1992;**101:**1174–1175.
15. Hnizdo E, Sylvain D, Lewis D, Pechter E, Kreiss K. New-onset asthma associated with exposure to 3-amino-5-mercapto-1,2,4-triazole. *J Occup Environ Med* 2004;**46:** 1246–1252.
16. Seed MJ, Cullinan P, Agius RM. Methods for the prediction of low-molecular-weight occupational respiratory sensitisers. *Curr Opin Allergy Clin Immunol* 2008;**8:**103–109.
17. Moore VC, Manney S, Vellore AD, Burge PS. Occupational asthma to gel flux containing dodecanedioic acid. *Allergy* 2009;**64:**1099–1107.
18. Cartmell J, Krstajic D, Leahy DE. Competitive Workflow: novel software architecture for automating drug design. *Curr Opin Drug Discov Devel* 2007;**10:**347–352.
19. Jonsdottir SO, Jorgensen FS, Brunak S. Prediction methods and databases within cheminformatics: emphasis on drugs and drug candidates. *Bioinformatics* 2005;**21:**2145–2160.