Connectionist Modelling of Asthma Incidence in New Zealand

Simon Hales[*], Qingqing Zhou[†], Simon Lewis[‡], and Martin Purvis[†]
[*]Department of Public Health, Wellington, New Zealand
[†]Information Science Department, University of Otago, Dunedin, New Zealand
[‡]Wellington School of Medicine, Wellington, New Zealand

Abstract

This paper describes an investigation of the patterns of self-reported asthma symptoms in relation to demographic and environmental factors in New Zealand. The subjects of the study consisted of 25,000 adults aged 20-44 who responded to a postal questionnaire. For each respondent, physical and social environmental conditions in the meshblock of residence were estimated using a Geographic Information System. The measured outcome was the 12-month prevalence of asthma. An artificial neural-network was constructed to model this outcome on the basis of the 13 environmental and demographic inputs, using a randomly selected sample of the data and tested on the remaining set of data. The modelling results indicated that there was different behaviour for the 20-25 and 26-44 age-groups, and separate neural-network models were constructed for each of these two age-groups. A set of inference rules were then extracted from each of the two neural networks. When applied to test data, the inference rules predicted the occurrence of asthma correctly in approximately 70% of the cases. Our approach may prove to be useful in simulating the effect of scenarios of environmental change on the occurrence of asthma.

1 Introduction

A national survey carried out during 1991-3 found substantial regional variation in asthma prevalence among New Zealand general electorates (Lewis, 1997). Subsequent analyses found evidence of weak associations between asthma symptoms and environmental factors (Salmond, 1998; Hales, 1998). Hales *et al.* (1998), observed substantial differences in asthma prevalence between general electorates and across quartiles of physical environmental factors. There has been considerable debate about the possible role of various environmental factors in explaining temporal and geographical patterns in asthma prevalence (Devalia,

Ruszniak *et al.* 1994; Martinez 1994; Seaton, Godden *et al.* 1994; Burr 1995; Newman-Taylor 1995; Strachan 1995; Balfe, Crane *et al.* 1996). Neural network models have been found to have better predictive accuracy than regression models in some data sets (Duh, 1998). In this study, we investigate the role of demographic, and environmental factors in patterns of adult asthma symptom prevalence in New Zealand by means of connectionist-based analysis.

2 Asthma prevalence in New Zealand

Asthma prevalence data were obtained from an extensive survey of New Zealand citizens that was conducted during 1991-93. The European Community Respiratory Health Survey (ECRHS) measured adult asthma symptoms and severity in a number of countries, using standardised methods. New Zealand participated in the ECRHS, initially involving surveys in Auckland, Hawkes Bay, Wellington and Christchurch in 1991-2. The survey was subsequently extended to cover the whole country in 1993. The methodology for the survey has been described in detail elsewhere (Burney, Luczynska *et al.* 1994). Briefly, a one page questionnaire was mailed to 31,470 people aged 20-44, chosen from the 1991 New Zealand electoral roll, sampling at least 1 in 40 from each electorate. Addresses of registered voters in the appropriate age range were obtained from the electoral office, and the questionnaire was sent to a random sample of these in each electorate, along with a letter explaining the purpose of the study. An attempt was made to telephone those people who had not responded after two reminders had been mailed to them. Respondents were asked to answer "yes" or "no" to seven questions relating to asthma, and to provide basic demographic details. The overall response was 82% (excluding ineligibles). The only modification made to the ECRHS questionnaire for the purposes of the New Zealand study was the addition of a question on ethnicity. Some relevant components of this survey are treated in the following paragraphs.

Asthma data

Asthma was defined according to the ECRHS definition (the proportion of subjects who reported one or more of the following: woken with shortness of breath in the past 12 months,

an attack of asthma in the past 12 months, or current asthma medication). Respondents were geocoded to meshblock level (an area of variable size depending upon population density) from the electoral roll.

Social deprivation

Social deprivation in the meshblock of residence was estimated for each respondent that could be geocoded using the NZDep91 index of deprivation (Crampton, Salmond *et al.* 1997).

Environmental data

Land use data and interpolated climate surfaces representing mean monthly average daily temperatures, rainfall, 9 a.m. humidity and solar radiation for the whole of New Zealand were obtained from Landcare Research (Leathwick, 1998). Individual estimates of exposure to each of these environmental variables were calculated for all questionnaire respondents that could be geocoded to an appropriate meshblock of residence. Environmental data and meshblock centroids were imported into in Arcview 'spatial analyst' (ESRI 1996). Estimates of each variable for each meshblock were generated using the 'summarise zones' function. Meshblock level data were merged with questionnaire data using Microsoft Excel 8.

The composite data set

The resulting accumulated data set consisted of 25092 instances, 21278 in Class 0 (no asthma) and 3814 in Class 1 (asthma). There were 10 attributes: 1) Age; 2) Ethnicity; 3) Humidity; 4) Land use; 5) Mean temperature; 6) NZDEP (social deprivation index); 7) PPT (rainfall); 8) Sex; 9) SRD (sunshine); 10) Wind speed. The range of the attributes was as shown in Table 1.

Table 1. Ranges of the data

Data category	Range
1) Age (years)	20-44
2) Ethnicity	(5:European; 4: Maori; 3: Pacific Islander; 2: Chinese; 1: Others)
3) Humidity	65.4% - 98.1%
4) Land use	52 classes – a quasi-ordered set with
5) Mean temperature	5 ^{o}C - 16 ^{o}C
6) NZDEP	1-10
7) PPT (precipitation)	30.9 mm/month - 678.8 mm/month
8) Sex	(0: Male; 1:Female)
9) SRD (sunshine)	12 hours/day - 15.4 hours/day
10) Wind speed	5.5 m/s - 24.9 m/s

3 The asthma connectionist model

Connectionist modelling was conducted using a three-layer feed-forward artificial neural network, with backpropagation training. Because the ethnicity component in the data set (item 2 in Table 1) represented a nominal value, there was a separate input ('yes' or 'no') for each of the five ethnic categories shown in Table 1. Thus there were 14 input nodes to the neural network and two output nodes (representing 'asthma' or 'no asthma'). The number of nodes in the intermediate layer of the network was then varied, and separate attempts were made to train the network in an effort to achieve the best results.

The training set consisted of two equal-sized, randomly selected sets (on the order of 1,000): one set selected from the group reporting 'asthma' and the other set from the group reporting 'no asthma'.

Although approximately 85% of all those surveyed reported having 'no asthma', our training sets were constructed such that there were equal subsets of those reporting 'asthma' and those reporting 'no asthma'. Under these circumstances satisfactory performance of a trained neural network would be to 'predict' the incidence of asthma in significantly greater than 50% of the

cases that are presented to it. It proved to be difficult to train these neural networks for satisfactory performance, however, probably reflecting the weak associations between asthma and the exposure factors (as noted in Hales *et al.*, 1998). After numerous trials, it was found that the best neural network training performance was achieved by partitioning the entire data set into two groups: one comprising the people with ages 20 to 25, and the other comprising people with ages 26 to 44. The characteristics of these two groups are shown in Table 2.

Table 2. Data set partitioned into two groups

Data Group	asthma	no asthma	total
Group 1 (age: 20-25)	866	4,237	5103
Group 2 (age: 26-44)	2948	17,041	19989

3.1 Asthma connectionist model for ages 20-25 (Group 1)

For training the neural network with Group1 data, 500 records were randomly selected from the 866 records reporting 'asthma' for training, and another 500 records were randomly selected from the 4,237 records reporting 'no asthma'. For the test data from Group 1, an additional set of 1700 records distinct from the training set was randomly selected: 850 with no asthma and 850 with asthma. Best results were achieved with a neural network having an architecture of 13 input units ('age' is now no longer an input), 65 hidden units, and 2 output units. Although training results were deemed to have a satisfactory accuracy, it was necessary to employ a large number of hidden nodes to achieve this result. These results are shown in Table 3.

Table 3. Neural net training for age-group 20-25 (Group 1)

Class	Training data			Test data		
	Total	Errors	Accuracy(%)	Total	Errors	Accuracy(%)
no asthma	500	61	87.8	850	244	71.3
asthma	500	35	93	850	184	78.4
Total	1000	96	90.4	1700	428	74.5

With this trained neural network, a set of inference rules were extracted using the following approach:

1. Train a multilayer feedforward neural network
2. Discretise the original input values and the hidden unit node activation values
3. Generate rules for transfer to each layer from the discretised values
4. Generate a composite rule set

In order to discretise the data, we used the Chi2 method (Liu & Setiono, 1995), which is oriented around the χ^2 statistic and consists of two phases. In the first phase each attribute *i* is associated with some significance level, say 0.5 at the outset. The data values of this attribute are sorted, and each value is considered to be resident of a (initially single-valued) interval. Then the χ^2 value is calculated for each pair of adjacent intervals. Starting with the lowest χ^2 value, adjacent intervals are merged until all pairs of intervals have χ^2 values greater than the χ^2 associated with the current significance level for the given attribute. This is done for each attribute. These merged intervals now represent a discretisation of the data set. With the reduced number of intervals, it is possible that there are now inconsistencies (two identical data elements associated with different output class values) in the data set. If the number of observed inconsistencies remains below a user set value, *, the above process is repeated with a decremented significance level for the attributes (and hence a larger tolerated χ^2 value). At the end of the first phase, the data set is discretised, and the number of data elements has been reduced

In the second phase of the Chi2 method, each attribute *i* is associated with an individual significance level, *sigLevel[i]*, and takes turns for merging. Consistency checking is conducted after each attribute's merging; if the consistency constraint is exceeded, attribute *i* will not participate in further merging. At the end of the second phase, no attributes can be further merged. If during this process an attribute has been merged to a single interval, then

it means that attribute is not useful for discrimination and can be dropped from further consideration. By this means feature selection can be achieved. Further details of the use of Chi2 can be found in (Purvis *et al.*, 1997; Purvis *et al.*, 1998).

When the input data and the hidden-layer activation values have been discretised, it is possible to generate rules from the input to the hidden-layr and from the hidden-layer to the ouput. For illustrative purposes we consider a simple case (Purvis *et al.*, 1998). Suppose we have a neural network with 13 input nodes and 3 hidden nodes and further suppose that Chi2 discretisation yields only a single discrete activation level for the first of the three hidden nodes and three discrete activation levels for each of the other two hidden nodes. This means that there are a total of 9 possible combinations of these hidden unit values. When the neural network is trained, however, it may be the case that only 5 of these combinations of values are actually observed in the training. If we consider `a(i,j)` to represent the `j`-th activation value of the `i`-th hidden-unit node, then we can number the 5 hidden-layer combinations as follows:

```
hidden-layer-combination #1= a(1,1),a(2,1),a(3,1)
hidden-layer-combination #2= a(1,1),a(2,2),a(3,1)
hidden-layer-combination #3= a(1,1),a(2,3),a(3,1)
hidden-layer-combination #4= a(1,1),a(2,3),a(3,2)
hidden-layer-combination #5= a(1,1),a(2,3),a(3,3)
```

The goal with respect to knowledge extraction is then to extract two sets of rules:

- one set of rules that describe which combinations of discretised inputs lead to certain hidden-layer combinations and
- another set of rules that describe how discrete hidden-layer combination values lead to discrete output values.

We perform rule extraction by using the X2R algorithm (Liu & Tan, 1995), which proceeds

in three steps:

1. Generate a rule to cover the most frequently occurring pattern. This is the shortest rule that can differentiate the pattern from patterns of other classes. Then remove this pattern from further consideration and iteratively repeat this step.
2. Generated rules are grouped in terms of their class labels.
3. For each rule cluster, remove redundant rules and drop more specific rules in favour of more general rules for the cluster.

This can result in rules like the following:

```
IF I(1,2) and I(2,1) and I(7,3) and I(13,2)
THEN hidden-layer-combination #1

IF I(2,3) and I(7,1)
THEN hidden-layer-combination #5
```

where `I(i,j)` is used to denote the `i`-th input unit taking its `j`-th cluster value.

We can also generate rules from the hidden-layer to the output, such as

```
IF hidden-layer-combination #1
THEN output-class = 1
```

When these rules are combined, we get rules like the following:

```
IF I(1,2) and I(2,1) and I(7,3) and I(13,2)
THEN output-class = 1
```

When there are a large number of hidden nodes, as was the case with Group 1, there will be a large number of rules generated, and so it was in this case. Using this approach, there were 253 rules generated. 48 of these rules did not involve ethnicity or sex, and two example derived rules are as follows:

Rule 1:

```
IF   Humidity > 81.5 and
     Land-use < 23 and
     Mean-temperature < 13.5 and
     7 < Social-index < 9 and
     Rainfall > 120.8 and
     12.4 < Wind-speed < 17.4
THEN 'asthma'
```

Rule 48:

```
IF   Humidity < 79.5 and
     Mean-temperature > 15 and
     Social-index < 4 and
     98.6 < Rainfall > 109.6 and
     11.4 < Wind-speed < 11.9
THEN 'asthma'
```

There were 36 additional rules generated for females and 27 rules for males (but which did not involve ethnicity). An example of such a rule is

Rule 85:

```
IF    sex is male and
      Humidity > 81.5 and
      Mean-temperature < 13.5 and
      7 < Social-index < 9
      Rainfall < 98.6 and
      Wind-speed < 11.4
THEN  'asthma'
```

The remaining rules included reference to specific ethnic groups. For example one of the 45 rules specific to Maoris was

```
Rule 112:
IF    ethnicity is Maori and
      Humidity > 81.5 and
      Social-index > 9 and
      109.6 < Rainfall < 118.3 and
      Wind-speed > 17.7
THEN  'asthma'
```

When these extracted rules were applied to the data for Group 1, the results were 75% accuracy for the training data and 63% accuracy for the test data. For a neural network with a smaller number of hidden nodes, there are fewer rules extracted, but the inference performance of the smaller rule set is worse. For example another neural network that was used with Group 1 data had only 17 hidden-layer nodes and produced an accuracy of 73% with the training data and 64% with the test data (compared with 90% and 75%, for the respective categories that was achieved by the neural network with 65 hidden-layer nodes presented in Table 1). The rule extraction process applied to this smaller neural network yielded fewer rules (140), but this rule-set had much poorer inference performance than the rules from the larger neural network.

3.2 Asthma connectionist model for ages 26-44 (Group 2)

For survey respondents in the 26-44 age-group, the best of the neural networks that were tested consisted of 13 input nodes, 50 hidden-layer nodes, and 2 output nodes. Since this was a larger data set than that of Group 1, the randomly chosen training data was also larger. The results achieved with this neural network after training are shown in Table 4.

Table 4. Neural net training for age-group 26-44 (Group 2)

Class	Training data			Test data		
	Total	Errors	Accuracy(%)	Total	Errors	Accuracy(%)
no asthma	1997	742	62.8	2897	1222	57.8
asthma	1998	597	70.1	2895	981	66.1
Total	3995	1339	66.5	5792	2203	62

Here the results were not as accurate as that achieved for Group 1 (Table 3), although as was the case for the Group 1 neural network, the performance of the model for predicting positive occurrences of asthma is better than it is for predicting negatives ('no asthma'). For this neural network the same rule extraction procedure was applied as was described in Section 3.1, and the rule extraction process yielded a set of 727 rules, of which 226 rules did not involve sex or ethnicity. Even though there were fewer hidden nodes for this neural network (which, by itself, would have the effect of reducing the number of extracted rules), the lower accuracy of the Group 2 neural network affected the discretisation process such that there was a larger set of extracted rules than that for Group 1. When these inference rules were used in conjunction with the Group data, the results were 70% accuracy for the training set and 62% for the test data set. These results are slightly better than the neural network from which they were extracted, and this seems to have been due to the manner in which the discretisation process set the interval boundaries for some of the data elements.

4 Discussion and conclusions

We have constructed neural network models to predict the likelihood that an individual has asthma, based on demographic and environmental attributes. The model for the Group 1 (age 20-25) was significantly better with respect to prediction than the model for the Group 2 (age 26-44). Because of the large number of hidden-layer nodes required to train these models, a question might be raised as to whether the networks have been overtrained and merely model the training data. However the networks were capable of successful prediction of asthma occurrence on novel test data, which would indicate that the networks have 'learned'

something about the association between asthma occurrence and the environmental and demographic parameters.

The extraction of inference rules from these trained neural networks resulted in a large number of rules. However, the rule sets demonstrated the capability to 'generalise' from the training data set to a novel set of test cases. This is not inconsistent with other research suggesting that the causes of asthma are complex. Certainly there are significant causal factors associated with the occurrence of asthma that are not among the input parameters of the collected data set. Nevertheless, both the neural network models and the extracted inference rule sets indicate that there are associations between environmental and demographic parameters and the occurrence of asthma in New Zealand. Many of the associations may reflect indirect causal mechanisms. We are continuing this series of experiments in an effort to produce more compact rule sets that still yield adequate prediction capability. By examining these rules, it may be possible to extract some useful information about the nature of asthma occurrence. In addition, by adjusting the input data to simulate scenarios of environmental change, it should be possible to use the model to predict the potential effects of such scenarios on asthma occurrence.

References

Balfe, D., Crane, J., and Beasley, R. (1996). "The worldwide increase in the prevalence of asthma in children and young adults." Continuing Medical Education 14: 433-42.

Burney, P., Luczynska, C. Chinn, S., and Jarvis, D. (1994). "The European Community Respiratory Health Survey." Eur Resp J 7: 954-60.

Burr, M. (1995). "Pollution: does it cause asthma?" Arch Dis Child 72: 377-87.

Crampton, P., Salmond, C., and Sutton, P. (1997). "NZDep91: a new index of deprivation." Social Policy Journal of New Zealand 9: 186-93.

Devalia, J., Ruszniak, C., and Davies, R. (1994). "Air pollution in the 1990s - cause of increased respiratory disease?" Respiratory Medicine 88: 241-4.

Duh, M., Walker, A., and Ayanian, J. (1988) Epidemiologic interpretation of artificial neural networks. American Journal of Epidemiolgy 147: 1112-1122.

ESRI (1996). Arcview Spatial Analyst. Redlands California USA, Environmental Systems Research Institute.

Hales, S., Lewis, S., Slater, T., Crane, J., and Pearce, N. (1998) Prevalence of adult asthma symptoms in relation to climate in New Zealand. Environmental Health Perspectives 106: (9).

Leathwick, J. and Stephens, R. (1998) Climate Surfaces for New Zealand. Wellington: Landcare Research.

Lewis, S., Hales, S., and Slater, T. (1997). "Geographical variation in the prevalence of asthma symptoms in New Zealand." New Zealand Med J 110: 286-9.

Liu, H. and Setiono, R. (1995) Chi2: Feature Selection and Discretization of Numeric Attributes. *Proceedings of the 7th International Conference on Tools with Artificial Intelligence*. Washington D.C., 388-391.

Liu, H. & Tan, S. T. (1995) X2R: A Fast Rule Generator. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (SMC'95), Vancouver, Canada, pp. 631-1635.

Martinez, F. (1994). "Role of viral infections in the inception of asthma and allergies during childhood: could they be protective?" Thorax 49: 1189-91.

Newman-Taylor, A. (1995). "Environmental determinants of asthma." Lancet 345: 296-9.

Purvis, M., Kasabov, N., Benwell, G., Zhou, Q., and Zhang, F. (1997) Neuro-fuzzy Methods for Environmental Modelling. *Proceedings of the Second International Symposium on Environmental Software Systems*, Chapman and Hall, London pp. 30-37.

Purvis, M., Zhou, Q., Cranefield, S., Ward, R., Raykov, R., and Jessberger, D. (1998) “Spatial Information Modelling and Analysis in a Distributed Environment” accepted for publication in *Environmental Modelling and Software*.

Salmond, C., Crampton, P., Hales, S., Lewis, S., and Pearce, N. (1998), Asthma and social deprivation. (submitted, 1998)

Seaton, A., Godden, D., and Brown, K. (1994). "Increase in asthma: a more toxic environment or a more susceptible population?" Thorax 49: 171-4.

Strachan, D. (1995). "Time trends in asthma and allergy: ten questions, fewer answers." Clin Exp Allergy 25: 791-4.