Interpretation of serial peak flow records for the diagnois of occupational asthma – a method based on pattern recognition technology.

Dr. Philip Bright MD, MRCP

Dr. Charles Pantin MD, FRCP

Mr Darren Newton

Dr. Sherwood Burge MD, FRCP

Correspondence to:

Dr. Philip Bright
Department of Chest Medicine
Birmingham Heartlands Hospital
Birmingham
B9 5SS

KEYWORDS

Version 2

11/11/98

Introduction

The diagnosis of occupational asthma is made from a suggestive history and the demonstration of changes in airway calibre when comparing days when the subject is at work to days when they are not.

Traditionally the monitoring of changes in airway calibre has been done by serial peak flow recordings which are then plotted before inspection by a reader expert in their interpretation. Such an analysis method is obviously subjective and liable to a lack of repeatability. The diagnosis of occupational asthma may be hindered by the scarcity of trained experts.

The visual interpretation of a serial PEF record is essentially pattern recognition and two classification methods based on a statistical technique - discriminant analysis, have been attempted. Both of these methods (Oasys-2 and Oasys-3){Gannon, Newton, et al. 1996 ID: GANNON1996A},{Bright & Burge 1995 ID: BRIGHT1995A} have good sensitivities and reasonable specificities when scoring whole records but the scoring of individual parts of serial peak flow records (complexes) shows discrepancies when compared to a human expert. This disagreement may lead to some whole records being wrongly classified.

Neural networks provide a radically different approach to the problem of classifying serial peak flow records for the presence of work related changes.

Aim: To develop a neural network based pattern recognition system trained to classify serial PEF records for the presence of work related changes in the PEF based on the classification of such records by a human expert.

To test the result in the new set of peak flow readings against a diagnosis made independently of the peak flow record.

Method

Patients attending a specialist clinic for the diagnosis of occupational asthma were asked to complete serial peak flow records as part of their clinical assessment. Patients were instructed to record the best of three blows as long as the best to next best difference was ≤ 10 L/min, every 2 hours from waking until going to sleep, on days at and away from work, for a period of three or more weeks. The times of waking, sleeping, starting and stopping work were also recorded.

Serial peak flow records were plotted as daily mean, maximum and minimum according to the method of Burge {Burge 1982 ID: 211}{Burge 1982 ID: 209}] (Figure 1). Three hundred serial PEF records, not previously used in the development of Oasys, were initially examined and quality control criteria applied.

Each record had to be a minimum of 21 days in duration and contain at least two periods of days at and away from work. On each day there had to be at least four peak flow readings{Gannon, Newton, et al. 1992 ID: 1937}. Records with evidence of respiratory tract infections were excluded from the study.

The 248 records that passed quality control checks were then divided randomly into 3 sets; a training set, a test set and an interrogation set.

A series of peak flow records were also collected to form Gold Standard sets and the above quality control criteria applied. Gold Standard positive (GS^+) records were those in which the diagnosis of occupational asthma was obtained by methods other than serial PEF records (a compatible history plus either specific bronchial challenge testing, a four-fold change in non-specific bronchial reactivity between times when the patient was exposed and not exposed or a positive RASTs to a relevant occupational antigen). Gold Standard negative (GS^-) were records in which the patients had been removed from exposure or were asymptomatic non-exposed workers. The GS^+ and GS^- sets were each divided randomly into two further groups ($GS1^+$, $GS2^+$, $GS1^-$, $GS2^-$)

The records in the training, test and interrogation sets were scored by a human expert Each record was examined by considering the record as a number of overlapping complexes (Figure 2). Work-Rest-Work complexes consist of a period of days at work, a period of days away from work then a further period of days at work. A Rest-Work-Rest complex consists of a period of days away from work, a period of days at work then a further period of days away from work. Each complex was awarded a probability from 0 for no evidence of a work-related change in the serial PEF plot to 100 for definite evidence of a work related change in the serial PEF plot. The probabilities from 0 to 100 were compressed into a score of 1≡probability 0, 2≡probabilities 1-49, 3≡probabilities 50-99 and 4≡probability 100 for use by the neural network program (NeuralShell 2{Ward Systems Group Inc. 1993 ID: WARDSYSTEMSGROU1993}),.

To enable the neural network program to manipulate the data, individual records were split into their constituent complexes, the PEF data for each complex was reduced to the daily maximum, mean and minimum. The work and rest periods that made up the complexes varied in duration and hence there was a range in length of the complexes. Data presented for analysis to the neural network needed to be of a consistent number of days. Each complex was therefore manipulated to form a two rest day, five work day, two rest day pattern for Rest-Work-Rest complexes or five work day, two rest day, five work day pattern for Work-Rest-Work complexes.

The neural network was then trained to interpret Work-Rest-Work and Rest-Work-Rest complexes from the training set of peak flow records, the network constantly testing itself against the test set of data. Once the best possible network for both Work-Rest-Work and Rest-Work-Rest complexes had been obtained the performance of the network compared to the expert was investigated by scoring the interrogation set of data

The developed neural networks were then used to score the complexes of records in $GS1^+$ and $GS1^-$. Scores of component complexes were combined to produce a score for the parent whole record. The scores of the component complexes were averaged with weightings of x2 given to complex scores of 1 and 4. The scores for whole records from $GS1^+$ and $GS1^-$ were used to determine cut-off points between records with a definite work effect, records with no work effect and records were no clear decision could be made. The higher of these cut-off points was used to calculate the sensitivity and specificity for the neural network (Oasys-NN) after scoring $GS2^+$ and $GS2^-$ records.

Results

The number of records in each of the data sets used is given in Table 1, together with the mean number of complexes in each record, the mean number of readings each day and the mean duration of records. Records from the training set tended to be shorter than those from the test and interrogation sets, this difference was also reflected in the number of complexes donated by each record from the three sets of records used to develop the neural network. The gold standard sets of records were all comparable except that $GS1^-$ records were slightly shorter than the other gold standard sets. The number of readings per day were equivalent for all sets.

The diurnal variation in the records from the individual sets is given in Table 1and is expressed both as the mean diurnal variation (% predicted) and the mean percentage of each record with a diurnal variation (% predicted) equal to or greater than 15%. The only set that is distinct from the others is GS1$^-$ which tended to have records with a much lower diurnal variation than any other set. GS1$^-$ consisted mainly of asymptomatic Post Office workers, while GS2$^-$ consisted of workers removed from exposure. The two gold standard positive sets tended to have the highest diurnal variation with GS2$^-$ being slightly lower.

Table 2, shows the suspected occupational sensitisers to which workers whose records make up the training, test and interrogation sets were exposed. Table 3and Table 4 give the method of diagnosis in patients making up GS1$^+$and GS2$^+$. For GS2$^+$, 12 patients had a diagnosis based on RAST only, 10 specific bronchial provocation tests only, and 1 non-specific bronchial reactivity changes only. 2 patients had a diagnosis based on RAST and specific bronchial provocation tests, 1 on RAST and non-specific bronchial reactivity changes and 5 on non-specific bronchial reactivity changes and specific bronchial provocation tests. In only one patient was the diagnosis based on positive results for all three diagnostic methods.

The agreement between the human expert and Oasys-NN for both the test and interrogation sets were reasonably close The weighted kappa score for the test set Work-Rest-Work complexes was 0.79 and 0.69 for Rest-Work-Rest complexes. The corresponding values for the interrogation sets were 0.83 and 0.74. For all complexes (Work-Rest-Work and Rest-Work-Rest) in the test set the weighted kappa was 0.74 and 0.79 for the interrogation set.

For whole records from the test set the mean (95% CI) difference in scores between the human expert and Oasys-NN was 0.0 (-0.11, 0.11) and limits of agreement 0.65 to 0.65, for the interrogation set the mean difference was 0.08 (0.01, 0.15) and limits of agreement 0.48 to 0.64 .

Of the inputs used by the neural network, the daily mean peak flows of the central parts of both Work-Rest-Work and Rest-Work-Rest complexes were most important in determining the classification of the complex.

Using GS1$^+$ and GS1$^-$ the sensitivity and specificity of the neural network for cut-off points for positive and negative records are given in Figure 3. Based on this graph, cut-off points of ≤ 2.0 for a negative record and ≥3.0 for positive records were chosen, with the area between indicating an uncertain diagnosis. For GS2$^+$, 25 records had a score ≥3, 6 records had scores in the range 2-3. No records had scores of ≤2. For GS2$^-$ records, 44 had scores ≤2, 12 had scores in the range 2-3 and no records had scores of ≥3. Based on a cut-off of ≥3 to indicate a definite positive record, scores below this indicating a negative record, the sensitivity for Oasys-NN was 80.6 and specificity 100%, at a cut-off point of 2.5 the sensitivity was 89.3 and the specificity 100% and at a cut-off of 2 the sensitivity was 100 and specificity 78.6%.

Discussion

With a sensitivity of 80.6% and a specificity of 100% at a cut-off between records showing a work related effect and those showing no work related effect of 3.0, Oasys-NN has been shown to be both sensitive and specific when classifying gold standard positive and negative records. In comparison to Oasys-2{Gannon, Newton, et al. 1996 ID: GANNON1996A}and Oasys-3{Bright & Burge 1995 ID:

BRIGHT1995A} whose sensitivities were 75% and 82% and specificities were both 94%, Oasys-NN would appear to be superior. The agreement between the human expert and Oasys-NN at a complex by complex level is also an improvement on that reported for Oasys-2 and Oasys-3. The weighted kappa scores for Oasys-2 and Oasys-3 for complexes in the test set used in developing Oasys-3 were 0.55 and 0.51 respectively compared with 0.74 for Oasys-NN scoring complexes from the test set used in its development and 0.79 for the interrogation set. This improvement for individual complexes is as important as the overall improvement in sensitivity and specificity as it reinforces the ability of Oasys-NN to match the human expert. For Oasys-2 and Oasys-3, the relatively poor comparison with the human expert on a complex by complex level makes mis-classification of whole records more likely. When comparing Oasys-NN with previous versions it is justified to choose a cut-point of 3.0 for Oasys-NN as this point is determined on the same criteria as were the cut-points for Oasys-2 and Oasys-3; maximum specificity with the best sensitivity subsequent to this.

A scoring system using three ranges of scores is more desirable allowing for scores in the grey area to indicate that their is doubt as to the nature of the record while defining records that fall into the other two ranges more certainly. The choice of the lower cut-off of 2.0 is fairly intuitive from the graph (Figure 3.) whereas the choice of the upper limit of 3.0 is less so. A logical cut-off would have been 2.6 which maximises specificity without too great a fall in specificity. Difficulties would arise in the acceptance of this analysis method by occupational health workers, who may well be loose confidence in the program if the it identifies a case as positive when it is not. An upper cut-off of ≥3.0 was chosen to allow a greater ‘safety margin’, even though the sensitivity was consequently lower than it would be at a cut-off of 2.6. In use records whose scores placed them in the grey area (scores between 2 and 3) would warrant further investigation with, for example, an examination of the quality of the records, an attempt to ensure that exposure to the agent was constant during the record and a repeat of the serial PEF recordings thereafter. The six GS2^{+} have been examined.

The diurnal variations (mean for record) for 4 of the records were below the 95%

confidence limit for the group as a whole, but only one had a diurnal variation below 15%. The scores awarded by OasysNN spanned the grey range from 2.10 to 2.69. In summary this group of records differs little from GS2$^+$ as a whole. If a cut-off score of 2.6 had been used between grey and positive records then four of these records would have been classified as having definite work effects. In 4 records it is possible that the exposures were intermittent based on the PEF changes in the plot. However, no change in exposure was reported by the workers. This may be inaccurate reporting, because the workers were unaware of exposure changes or because some other factors were involved, for instance, unreported changes in treatment. Two records had small but definite changes. Although these tended to be reported as '3', i.e., probable work effect, these scores are not emphasised by the weighting system. The significance of small changes has not been fully explored.

Although excellent sensitivity and specificity were achieved the problem of gold standard positive records tending to be from patients with more obvious PEF changes remains except from those diagnosed from a history and positive RAST, when no prior degree of pulmonary function test change is required. However, the ability of Oasys-NN to classify records well is supported by the closeness of agreement between the program and the expert both for individual complexes and whole records for the interrogation set.

Defining gold standard negative records can be difficult. In the work of Gannon{Gannon, Newton, et al. 1996 ID: GANNON1996A} asymptomatic post-office workers and those removed from exposure were used. There may be instances when such subjects may show definite work related changes in their PEF record either because of incomplete removal from exposure or because they have asymptomatic exposure to an unidentified sensitiser (the post-office study was carried out following the identification of occupational asthma in one worker). In determining the sensitivity and specificity of Oasys-NN the novel approach of generating gold standard records by using records of subjects completely away from work has been taken. Such records could only display an apparent work related pattern by chance or if 'pseudo-occupational asthma' was present. The later situation arises if a subject records the first PEF reading of the day later after waking at weekends than weekdays. As the first reading of the day is often the lowest the normal diurnal rise in PEF would ensure that at weekends the daily minimum and mean PEFs were higher than during the week.

This study has several advantages over previous work. Most studies rely on data from workers in one industry exposed to one respiratory sensitiser such that changes in the serial PEF record are likely to be stereotyped. In contrast, the records in this study have been culled from workers in many occupations exposed to many agents. For those studies based on single agents the ability of the serial PEF record interpretation techniques to classify records from workers exposed to other agents is uncertain and may be limited, whereas, for Oasys-NN, the system should generalise well. Strict quality control criteria have been applied and the characteristics of the data entered into the study investigated. While this means that there may be questions as to the validity of using Oasys-NN to analyse records of lesser quality, previous work has often failed completely to address this issue. The quality control criteria also establish minimum data standards that can be used when collecting data for use with the system.

Neural networks produce much tighter data fits than other techniques, excelling at classification problems where pattern recognition is important and precise computational answers are not required. Determining before hand the interdependence of indices is not important as neural networks consider every interaction. Neural networks are also able to cope with 'noisy' or slightly incorrect data without loss of classifying ability.

Neural networks are not suited to all situations and are difficult to interrogate as to the exact structure they use to produce the classification; as such, they are often viewed as 'black boxes'. This is in part unfair. The input indices with most influence can be determined by summating all the neurone weights that lead a particular input. For Oasys-NN the daily means were the most important inputs and were the ones that the human expert instinctively used the most in determining classification of a complex. It is possible to 'over-fit' a network to a set of data, so that while the network classifies the training patterns well it fails to generalise to novel patterns. This is usually the fault of allowing training to progress too far. In the neural network development program used in this study, by constantly testing the neural network at its ability to classify a set of data (the test set), the problem of poor generalisation due to over-fitting of the data should not arise. However, to be absolutely certain, Oasys-NN was then tested on a further set of data (interrogation set). Such rigorous testing is only possible because of the large amounts of data available and should guarantee that the network classifies novel patterns reasonably well.

The main potential problem with the methodology in this study is the degree of manipulation needed to produce data in a format that the neural network could use. In generating essentially artificial data based on the original complex data the true patterns may have been corrupted. However, as the method of manipulation is constant and as the network trains to match these manipulated patterns to the expert score, which was based on the original patterns, the manipulation should not interfere with the classification of novel patterns.

The use of a neural network based system to classify serial PEF records into those that do and do not show a work related effect and also into a group where no decision can be made has provided a sensitive and specific tool, the performance of which approaches, if not equals that of a human expert. In doing so, one of the main problems in diagnosing occupational asthma, that of access to an expert opinion on the serial PEF record has been addressed.

Oasys-NN should now supersede other methods of interpreting serial PEF records

for the presence of work related effects.

Table 1. Details of sets of records used.

	Number of records	Mean (95% CI) number of complexes per record	Mean (95% CI) number of readings per day	Mean (95% CI) duration of record in days	Mean percentage of each record with diurnal variation (% predicted) ≥15% (95% CI)	Mean diurnal variation (% predicted), (95% CI)
Training Set	172	4.7 (4.3, 5.1)	7.7 (7.4, 7.9)	26 (24.7, 27.3)	55.5 (50.4, 60.6)	21.2 (19.5, 22.9)
Test Set	33	12.9 (11.2, 14.6)	9.2 (8.2, 10.1)	61.7 (55.2, 68.2)	48.9 (40.9, 57.0)	18.7 (15.8, 21.5)
Interrogation	43	7.5 (6.6, 8.3)	7.9 (7.5, 8.4)	43.5 (40.5, 46.5)	56.1 (43.2, 69.1)	20.5 (16.9, 24.1)
GS1$^+$	56	6.5 (5.6, 7.3)	7.5 (7.1, 7.9)	32.1 (29.3, 34.9)	69.0 (61.2, 76.9)	25.9 (22.1, 29.6)
GS1$^-$	46	4.5 (4.0, 5.0)	8.4 (7.8, 8.9)	22.1 (20.3, 23.9)	11.7 (6.2, 17.3)	8.7 (7.4, 10.1)
GS2$^+$	32	6.7 (5.3, 8.0)	7.7 (7.0, 8.4)	35.7 (29.8, 41.6)	69.9 (59.1, 80.6)	27.4 (22.6, 32.1)
GS2$^-$	56	7.7 (6.8, 8.5)	8.2 (7.7, 8.8)	36.5 (32.9, 40.0)	55.0 (46.5, 63.5)	21.5 (18.5, 24.6)

- Table 2. Suspected occupational sensitisers for workers making up the training and interrogation sets

Sensitiser	Training Set	Training Set (%)	Test Set	Test Set (%)	Interrogation Set	Interrogation Set (%)
Colophony	44	26	8	24	10	23
Flour	10	6	2	6	3	7
Isocyanate	32	19	8	24	6	14
Oil Mist	19	11	5	15	9	21
Shrink Wrap	12	7	5	15	1	2
Unknown	43	25	4	12	12	28
Other	12	7	1	3	2	5

- Table 3. Exposures to occupational agents from subjects in gold standard positive set `(From Gannon et al(67))

Number of records	127
Isocyanate exposure	7
Oil mist exposure	6
Metal Exposure (Cr, Ni, Co)	5
Flour exposure	3
Colophony exposure	7
Epoxy resin exposure	6
Glutaraldehyde exposure	4
Wood dust exposure	0
Post office dust exposure	57
Other exposures	32
Method of independent diagnosis	
Specific challenge	31
Bronchial reactivity	13
Positive IgE RAST	13
Asymptomatic post office worker	58
Other	12

- Table 4. Details of the diagnosis of patients in gold standard positive set 2

Sensitiser	RAST positive	4x change in non-specific bronchial reactivity	Specific bronchial reactivity positive
Flour	Yes		
Glutaraldehyde			Yes
Glutaraldehyde		Yes	Yes
Glutaraldehyde		Yes	Yes
Flour	Yes		
Flour	Yes		
Flour	Yes		
Glutaraldehyde			Yes
Glutaraldehyde	Yes		Yes
Latex	Yes	Yes	Yes
Isocyanate			Yes
Isocyanate			Yes
Isocyanate	Yes		
Colophony			Yes
Colophony			Yes
Isocyanate	Yes		
Shrink wrap			Yes
Flour	Yes		
Isocyanate	Yes		
Isocyanate	Yes		
Flour	Yes		
Flour	Yes		
Isocyanate		Yes	
Colophony			Yes
Colophony			Yes
Colophony			Yes
Cutting oil		Yes	Yes
Flour	Yes		
Isocyanate	Yes	Yes	
Cutting oil		Yes	Yes
Shrink wrap		Yes	Yes
Shrink wrap		Yes	Yes

- Figure 1.

Diagram of daily maximum, mean and minimum serial peak flow plot.

-

Figure 2.

Diagram to show derivation of complexes from serial peak flow plots

- Figure 3.

Sensitivities and specificities over a range of cut-off points between records with and without work effects