Processing and representation of meta-data for sleep apnea diagnosis with an artificial intelligence approach

David Nettleton a,*, Joaquín Muñiz b

a Consulting Group—Business Intelligence, IBM Global Services, Barcelona, Spain
b Sleep Centre, Hospital de la Santísima Trinidad, Salamanca, Spain

Abstract

In this article, we revise and try to resolve some of the problems inherent in questionnaire screening of sleep apnea cases and apnea diagnosis based on attributes which are relevant and reliable. We present a way of learning information about the relevance of the data, comparing this with the definition of the information by the medical expert. We generate a predictive data model using a data aggregation operator which takes relevance and reliability information about the data into account to produce a diagnosis for each case. We also introduce a grade of membership for each question response which allows the patient to indicate a level of confidence or doubt in their own judgement. The method is tested with data collected from patients in a Sleep Clinic using questionnaires specially designed for the study. Other artificial intelligence predictive modeling algorithms are also tested on the same data and their predictive accuracy compared to that of the aggregation operator. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Questionnaire screening; Sleep apnea diagnosis; Relevance and reliability weights; Aggregation; Grade of membership; Categorical and scalar representation

1. Introduction

Screening of Apnea cases often involves a questionnaire which is filled in by the patient, but the success of this method in diagnosing apnea is often inhibited by the introduction of erroneous or inconsistent data by the patient. In this article we discuss some of these problems, citing possible solutions, and we detail a way of learning meta-data (data about data) with respect to the relevance and reliability of the questions in the questionaire. Another central aspect of the work is the use of a data aggregation algorithm (WOWA) to model the data and give a diagnosis for each patient based on the clinical data and questionnaire responses, incorporating relevance, reliability and membership grade information. The relevance information is defined as a weighting vector, which is learned from the data by a genetic algorithm,
and defines the relevance of each variable to the output (diagnosis). We also compare the use of relevance information learned by the genetic algorithm, to relevance information assigned manually by the medical expert. The reliability information is also defined as a weighting vector, and is always assigned manually by the medical expert and the data analysis expert for each variable, and interpreted for each real data value. It defines the reliability of each variable data response. The membership grade is read off a scale for each linguistic label and interpreted by a predefined function (membership) curve. The membership curve is defined manually by the data analyst and then revised by the medical expert. It defines the grade of membership of a response to one or more linguistic labels, such as ‘always’, ‘never’, and so on.

The paper is structured as follows: in Section 2, a clinical description of the sleep apnea syndrome and the diagnosis process is given; in Section 3 we explain the different artificial intelligence techniques used in data capture and processing (meta-data, fuzzy-scalar representation, WOWA data aggregation, and assignment of the meta-data (learning of meta-data with genetic algorithms and manual assignment by the medical expert); in Section 4, we discuss some of the problems with standard questionnaires and give examples for improvements which can help to eliminate these shortcomings; in Section 5, we outline the test data used for testing the methods; Section 6 gives a summary and analysis of the questionnaire responses comparing the scalar (fuzzy) representation with the categorical form, and summarizes results of processing the test data with the methods detailed in Section 3. We also compare the results using weights assigned manually by the medical expert with weights learned automatically by the genetic algorithm; Section 7 summarizes and gives some conclusions for the present work.

With reference to Fig. 1, the data processing scheme allows us to compare categorical and scalar questionnaires for diagnostic accuracy, and compare results when the ‘relevance’ weights are assigned by the medical expert (based on personal knowledge and the literature) or learned by the genetic algorithm. The ‘reliability’ weights are always assigned by the medical expert (based on personal experience and a knowledge of the specific case data captured in his Clinic). This is because it was considered that ‘reliability’ would be more difficult to learn statistically, whereas ‘relevance’ is more akin to correlation analysis. Note that the scalar questionnaire makes use of a manually defined membership function to interpret the patient’s response. The ‘reliability’ and ‘relevance’ values, on the other hand, are automatically interpolated into curves as part of the aggregation operator itself.

2. The sleep apnea syndrome and its diagnosis

The Sleep Apnea Syndrome is a frequent problem, which to a greater or lesser extent affects between 2 and 4% of the adult population in the developed countries [1,2]. It is characterized by complete (apnea) or partial (hypopnea) interruption of respiration during sleep. The presence of this syndrome has been associated with excessive somnolence,
with consequences such as traffic accidents and the reduction in quality of life and professional development [3]. It has also been linked to cardiovascular illnesses, there being a greater prevalence of hypertension, cardiac arrhythmia’s, cardiopathic ischemia and cerebral-vascular accidents (stroke) in these patients.

It has also been demonstrated that the presence of these symptoms is correlated with RDI (defined as the number of apneas + hypopneas per hour). Patients with low RDI’s, that is, less than 5 apneas, do not tend to have clinical consequences. Light cases, between 5 and 20, have slight consequences while moderate cases, between 20 and 40 usually show clinical manifestations. Severe cases, with an index above 40, show the most evident symptoms and present an increase in illnesses and death [4,5].

The diagnosis of Sleep Apnea Syndrome, and the categorization of its seriousness (light, moderate and severe) is achieved by the evaluation of a combination of clinical manifestations and data derived from a polysomnogram. The polysomnogram consists of a continuous recording, during nighttime, of numerous physiological variables, including electroencephalogram, electrooculogram, electromyogram, leg movement, oral-nasal airflow, snoring, thoracic and abdominal respiratory effort, electrocardiogram, body position and hemoglobin oxygen saturation. Other biological signs can also be used if considered necessary.

Due to the high cost of this type of clinical study, and the shortage of adequate centers, a series of more limited tests have been devised, which can be used for ‘screening’ in diagnosis. In general, the tests consist of a reduced number of variables (for example, only the pulsoximetry), which allow non-supervised studies to be made in the patients own home [6].

One of the most interesting tools available for diagnosis, due to its simplicity and low cost, are self-administered or supervised questionnaires. Having identified a set of variables with high predictive value for sleep apnea syndrome, diverse questionnaires have been developed, with combinations of questions and clinical variables. Unfortunately, this method has not found great acceptance in clinical use, due to its low predictive accuracy and the numerous false negative and positive diagnosis that it produces [7].

With the objective of improving this tool, we have designed a study where the patient was given a general sleep questionnaire, which permits a double evaluation in a scalar and a categorical form for each question, in order to see if the scalar form extracts a greater information from the patient and thus produces a greater correlation with the RDI. For diagnosis we have used an aggregation operator called WOWA, which is described in the following Section 3. This operator receives as input the available clinical data for each case and produces as output a diagnosis which can be positive or negative.

3. Artificial intelligence techniques used for data representation and processing

In this section we describe four techniques from the artificial intelligence field which have been applied in this work, namely: meta-data in the form of ‘relevance’ weightings for variables and ‘reliability’ weightings for data values; fuzzy (scalar) representation for recording of questionnaire responses; an aggregation method called WOWA for diagnosis, which uses as input the ‘relevance’ and ‘reliability’ meta-data; and the use of a genetic algorithm to learn the relevance weights.
3.1. Types of meta-data used to improve diagnosis

We consider three types of meta-data which relate to the questionnaire data: Reliability information—this is defined for each variable-response. The reliability weights are always assigned manually by the medical expert and the data analysis expert. Relevance information—this is defined for each variable. The relevance weights are assigned by two methods: in the first method they are learnt from the data by a genetic (evolutive) algorithm, as described later in Section 3.4; in the second method the relevance weights are assigned by the medical expert. This permits a comparison of the diagnostic precision of the two assignment methods. Grade of membership information—this is collected for each individual question variable, and is a number between 0 and 1, where 1 implies total membership to a linguistic label such as ‘always’ or ‘never’, and 0 implies no membership.

3.2. Data representation—scalar and categorical

3.2.1. Fuzzy representation

We designed a symmetrical and steep gradient membership function which can be overlaid on the scale of each question to read off the grade of membership to each linguistic label, the labels being {never, rarely, sometimes, frequently and always}. The function is designed to give a comparable response between questions and labels, and a steep change in the midrange which makes for greater sensitivity.

The patient draws a cross on the continuous scale to indicate his/her response to the question. For example, in the scalar questionnaire, a question to the patient would appear as:

<table>
<thead>
<tr>
<th>never</th>
<th>rarely</th>
<th>sometimes</th>
<th>frequently</th>
<th>always</th>
</tr>
</thead>
</table>

The fuzzy response is stored as a quintuple, with a membership grade for each linguistic label. For example the response to S5 (above) would be stored as: {0:0.0, 1:0.3; 2:0.7; 3:0.0; 4:0.0}. This indicates that only linguistic labels ‘rarely’ and ‘sometimes’ have non-zero membership values, being 0.3 and 0.7 respectively. We apply the weighted mean aggregator to this in order to produce one data value for the response. In the example this would be (0.0 × 0.0 + 1.0 × 0.3 + 2.0 × 0.7 + 3.0 × 0.0 + 4.0 × 0.0) giving 0.3 + 1.4 = 1.7. Note that we can convert to categorical if we so desire by rounding this value. In order to read the fuzzy scalar values, a transparent sheet with the function curves of Fig. 2 (above) superimposed is overlaid on each response line and the membership grade is read off on the y-axis.

3.2.2. Categorical representation

In the categorical form the patient must indicate one and only one category as his/her response. Thus:

<table>
<thead>
<tr>
<th>never</th>
<th>rarely</th>
<th>sometimes</th>
<th>frequently</th>
<th>always</th>
</tr>
</thead>
</table>

S5. Do you fall asleep while driving on the motorway?

Fig. 2. Fuzzy membership function use to interpret question scalar response.

S5. Do you fall asleep while driving on the motorway?
3.3. Data aggregation techniques — WOWA

Data fusion and aggregation operators are used in different fields, and in artificial intelligence, to fuse information supplied by different information sources. In recent years, several methods and techniques have been developed to deal with information represented under different forms: numerical, ordinal, etc. For example, in the particular case of the numerical setting, there exist, among others, the weighted mean [8], the Ordered Weighted Average (OWA) operator [8] and the Weighted OWA (WOWA) [9]. Usually, these aggregation operators are parametric with respect to a set of parameters, for example, one or two weighting vectors in the case of the weighted mean, OWA and WOWA operators. Different aggregation methods contrast different weighting factors; WM weights the data values, OWA weights the variables, and WOWA weights both the variables and the data values.

In this paper, we consider a data aggregation method based on the use of two weighting vectors to ‘bias’ the data, where bias is understood as the act of inhibiting or enhancing the contribution of a given variable or data item to the overall aggregated result. If we think a variable is more relevant to the overall result, we enhance its contribution, and if we think a variable is less relevant then we inhibit its contribution. Thus, in our particular application, the two weighting vectors used by WOWA are interpreted as representing the relevance of each variable and the reliability of each data item. We use two aggregation methods: weighted ordered weighted average (WOWA) and weighted mean (WM). We use WM to aggregate the membership grades with the ordered responses for the scalar data as explained in Section 3.2, while WOWA [9] is used to aggregate the data responses together with the reliability and relevance weights. Note that the aggregation of the membership grades is a distinct process to the posterior WOWA aggregation which produces the final aggregated result.

3.3.1. The WOWA operator

Torra in [9] defined the Weighted OWA operator (WOWA), which aggregates the input data into a single output value, using the clinical data itself, together with two weight vectors: \( \rho \) corresponding to the ‘relevance’ of the variables, and \( \omega \) corresponding to the ‘reliability’ of the data values. One of the difficulties in using aggregation operators is the initial fixing of the associated parameters, for example the relevance weights \( \rho \) of each information source.

Note that we distinguish the weighting of the reliability of a given questionnaire response from the reading of the membership grade value of the responses in each case as detailed in Section 3.2.1. We can say that the membership grade is reflecting the qualitative information provided by the patient, whereas the \( \omega \) weighting of the responses reflects the medical experts knowledge of what responses are most expected for each question, see [10,11] for previous studies of these aspects. Considering the weights as data points, WOWA uses the interpolation method of Chen and Otto [12], to create a continuous function curve which can be used to weight all the values of each variable. We now give some formal definitions for the aggregation operators.

Definition 1. A vector \( v = [v_1, v_2, \ldots, v_n] \) is a *weighting vector* of dimension \( n \) if and only if \( v_i \in [0, 1] \), \( \sum v_i = 1 \).

Definition 2 [9]. Let \( p \) be a weighting vector of dimension \( n \), then a mapping \( WM: \mathbb{K}^n \rightarrow \mathbb{K} \) is a weighted mean of dimension \( n \) if \( WM_p (a_1, \ldots, a_n) = \sum p_i a_i \).
Definition 3 [8]. Let $\mathbf{w}$ be a weighting vector of dimension $n$, then a mapping $\text{OWA}_w: \mathbb{R}^n \rightarrow \mathbb{R}$ is an Ordered Weighted Averaging (OWA) operator of dimension $n$ if $\text{OWA}_w (a_1, ..., a_n) = \sum_{i=1}^n \omega_i a_{\sigma(i)}$ where $\{\sigma(1), ..., \sigma(n)\}$ is a permutation of $\{1, ..., n\}$ such that $a_{\sigma(i)} = a_{(i-1)}a_{\sigma(i)}$ for all $i=2, ..., n$. (i.e., $a_{\sigma(i)}$ is the $i$-th largest element in the collection $a_1, ..., a_n$).

Definition 4. Let $\mathbf{p}$ and $\mathbf{w}$ be two weighting vectors of dimension $n$, then a mapping $\text{WOWA}: \mathbb{R}^n \rightarrow \mathbb{R}$ is a Weighted Ordered Weighted Averaging (WOWA) operator of dimension $n$ if $\text{WOWA}_{(p,w)} (a_1, ..., a_n) = \sum_{i=1}^n \omega_i p_{\sigma(i)} a_{\sigma(i)}$, where $\{\sigma(1), ..., \sigma(n)\}$ is a permutation of $\{1, ..., n\}$ such that $a_{\sigma(i)} = a_{(i-1)}a_{\sigma(i)}$ for all $i=2, ..., n$. (i.e., $a_{\sigma(i)}$ is the $i$-th largest element in the collection $a_1, ..., a_n$), and the weight $\omega_i$ is defined as $\omega_i = w^* (p_{\sigma(i)}) - w^* \sum_{1}^{(i-1)} p_{\sigma(i)}$ with $w^*$ a monotone increasing function that interpolates the points $(0,0), (1,0), ..., (n-1,0)$ and $(n,0)$. The function $w^*$ is required to be a straight line when the points can be interpolated in this way.

To diagnose a patient, WOWA is called thus:

$$\sum_{j=1}^n (A_i, V^\rho, V^{\omega_j}),$$

where $A_i$ is the data vector for patient $i$, $V^\rho$ is the ‘relevance’ weight vector for all variables, and $V^{\omega_j}$ is the ‘reliability’ weight vector which acts on the data values corresponding to variable $j$.

### 3.4. Meta data: ‘relevance’ and ‘reliability’ weight assignment

In this section we describe how a genetic algorithm can be used to learn the relevance weights $\rho$, from the original data. The second set of weights, the reliability weights $\omega$, are always assigned manually by the medical and data analysis expert, as described in the second part. The relevance weights are also assigned by the medical expert, in order to contrast the results with the automatic assignment method.

#### 3.4.1. Learning the ‘relevance’ weights from historical case data using an evolutive program (genetic algorithm)

If $O_p$ is the diagnosis predicted by the aggregation function WOWA, and $O_r$ is the normalized RDI value, that is the real diagnosis, then the objective is to minimize the difference between the sum of the $O_p$ and the sum of the $O_r$ for all patient cases. That is:

$$\text{Min} \left( \sum O_p - \sum O_r \right).$$

We can use genetic algorithm techniques to learn the $\rho$ (relevance) weighting factors from historical data values. Genetic algorithms mimic the natural selection process in which only the fittest survive from the current generation to go on to the next. A genetic algorithm has a set of input and output data (examples), a set of modifiable variables (in this case the weighting factors), a set of constraints (in this case the sum of the weighting factors must be equal to 1), and an objective function, which in our case is to minimize the difference between the predicted diagnosis and the real diagnosis. We wish to find the weighting factors which best approximate the input and output data, while minimizing the objective function.

**Example:**

Input (data): $I = \{5, 6, 4, 6, 1, 5, 2, 5, 2, 2, 1, 2, 2, 5, 9, 7, 2, 2, 8\}$, one value for each variable.

Input (relevance weights to be learned): initial values of $\rho = \{0.90, 0.20, 0.85, 0.25, 0.50, 0.90, 0.62, 0.90, 0.63, 0.68, 0.55, 0.67, 0.61, 0.93, 0.74, 0.63, 0.64, 0.27, 0.94\}$, one value for each variable.
Output: $O_p = [0,1]$ (predicted diagnosis value) $O_r = [0,1]$ (real diagnosis value) and, $O_c = \text{RDI}/\text{max (RDI)}$. The genetic algorithm uses a set of modifiable values, being $n$ vectors of $\rho$ (initial weights or initial population), and the value of the objective function $O_c$ for each corresponding row $n$ of data. The inner loop of the algorithm (Fig. 3) executes for a given set of weights $\rho_i$ and for all cases, to calculate the $O_p$. The outer loop repeats for each weight vector $\rho_i$ in the population, and the vectors are ranked in terms of the precision of the result, which is the sum of the differences between the calculated values $O_p$ and the true values $O_r$. In the following example of an execution of the inner loop, $O_r = 0.78$ indicates a fairly positive result, 0.00 being totally negative and 1.00 being totally positive.

For $\rho = \{0.90, 0.20, \ldots, 0.27, 0.94\}$ and $\rho_i = \{2, 8\}, O_p = 0.40$, $O_r = 0.36$

$I_2 = \{4, 7, \ldots, 3, 2\}, O_p = 0.19$, $O_r = 0.26$

$I_i = \{4, 8, \ldots, 3, 1\}, O_p = 0.82$, $O_r = 0.78$

Set of constraints:

(i) Values of $\rho$ between 0 and 1, with precision of 2 decimal points; (ii) sum of $\rho$ values equal to 1; (iii) values of $\rho$ must be normalized.

In Fig. 3 we see the ‘evaluate’ procedure which simply goes through all the individuals in the current population and assigns a ‘fitness’ score to each. The fitness score for each individual is calculated by executing the function which calculates the diagnosis (WOWA aggregator) for each of the patient cases (1..j) and with the $\rho$ weight vector contained in the chromosome of individual $i$.

3.4.2. Manual assignment of the ‘relevance’ and ‘reliability’ weights

Assignment of $\omega$ ‘reliability’ weights: the $\omega$ weight vectors depend on the real values of the data and are assigned in accordance with the range values defined by the expert for each variable. The reliability weight vector imposes an emphasis on a part of the range of values of a variable, for example, on the high values, low values or mid-range values.

The relevance and reliability weights are assigned on the one hand by the medical expert, based on current clinical literature in the Apnea diagnosis field, on his own knowledge and experience, and taking into account

---

**Procedure evaluate**

```
begin
  for all (genetic) individuals do begin
    read $\rho$ weights from current individual i’s chromosome
    for all patient data cases do begin
      read (patient data input row j)
      calculate $\omega$ weights from expert assigned ranges and each data value
      real_diagnosis ← read (real_diagnosis for this patient)
      wowa_diagnosis ← wowa (weights, data)
      local_distance ← (real_diagnosis – wowa_diagnosis)$^2$
      total_distance ← total_distance + local_distance
    end
    individual(i).aptitude ← total_distance
    total_distance ← 0
  end
```

Fig. 3. The basic structure of the evaluation routine.
the type of patients (the mix) which exists in the Salamanca Sleep Clinic with whom we collaborated for the study. In the previous section we have seen a method for ‘automatically’ generating the relevance weights by applying a ‘genetic’ learning algorithm to historical case data. In Section 6 the results of diagnosis using the WOWA aggregation operator with ‘automatically’ assigned ‘relevance’ weights is compared for precision with the results of diagnosis using the WOWA aggregation operator with ‘relevance’ weights assigned ‘manually’ by the medical expert.

4. Inherent problems of standard questionnaires and proposed solutions

The purpose of the questionnaire is to provide an information profile of the patient which allows a pre-diagnosis of his/her condition. This acts as a ‘screening’ which avoids patients being admitted into the sleep center for expensive and time consuming testing, when they have a low probability of suffering from Apnea Syndrome, or have some other pathology.

The questionnaire we have used consists of two main sections: the first records clinical data, with 15 key clinical variables: age, sex, presence of a partner, profession, work hours, education level, weight, height, neck circumference, BMI (body mass index), blood pressure, alcohol intake, cigarette intake, auto-evaluation of most important symptoms, other illnesses; the second section consists of 41 questions to which the patient responds on a five point scale {never, rarely, sometimes, frequently, always}. The questions are divided in 3 subsections: 15 general sleep questions (coded as G1 to G15), 16 respiratory related questions (coded as R1 to R16) and 9 somnolence related questions (coded as S1 to S9). Based on this information, the doctor then gives a clinical evaluation: healthy; simple snorer; doubtful; typical apnea; other illness. We interpret this as: typical apnea; no apnea, with the corresponding grade of membership.

One of the fundamental problems with the questionnaire responses is that in the general sleep and respiratory related questions, there are several key questions which rely on the bed partner as a witness. Of course if there is no bed partner, or the bed partner does not know, this eliminates some key information for the diagnosis. To improve this situation, in the case of there being a bed partner, we propose that s/he fills in the same questions separately in a different questionnaire. The responses can then be cross checked for contradictions and inconsistencies between the bed partner and the patient.

At the Sleep Center of the Hospital de la Santisima Trinidad of Salamanca we have obtained 71 questionnaires filled in by sleep patients. Each patient filled in two versions of the questionnaire — one with categorical responses and the second with fuzzy scalar responses. Each patient was previously briefed as to how to fill in the questionnaires. In practice, the patients were from all types of backgrounds, educational and cultural levels. Sometimes there were errors in how the patient responded to the categorical and scalar response representations. One typical error is that the patient responds to the scalar representation as if it were categorical, placing a cross exactly on the label point in each case. Thus there was no ponderance of a grade of membership by the patient. The lesson we have learnt from this is to dedicate more time to explaining to each patient the importance of thinking about the scalar response in order that they can appreciate our objectives in doing this. For example, the added subtlety of placing a cross on the scale, say, two thirds of the way between frequently
and always, but closest to always. Of course, as the subjects are taken randomly from the general public, this is not an easy task.

Finally we have a problem of data—having sufficient cases with which to test our method. Take into account that for many machine learning techniques for each N input variables we need N × 10 cases. The data should also ideally represent a homogeneous group of the population, such as professional males between 45 and 65 years of age with medium education level, living in the same geographical area, without secondary ailments. This is not our case, but instead we have real data with which any clinic has to deal with every day, and our objective is that our data processing methods give useful and acceptable results from it.

The patients were chosen at random for the study by the Sleep Pathologies Center, and studied with respect to diverse problems: insomnia, somnolence, snoring, apneas, body movement during sleep, nocturnal choking, etc. ... These patients filled in the questionnaire and were also given a complete night-time polysomnogram, or a supervised night-time cardio-respiratory polygram, in the Sleep Unit. The following variables were recorded: oral-nasal airflow, snoring, thoracic and abdominal respiratory effort, body position, actimetry, electrocardiogram, pulse and oxygen saturation in hemoglobin. The RDI has been determined for all patients, and this value was used to compare the predictive accuracy of the different types of questionnaire.

5. Test data—selected variables

The questionnaire is designed to detect diverse sleep pathologies. Thus the medical expert has selected a subset of variables with highest correlation specifically with Apnea diagnosis.

The variables selected by the medical expert are as follows: age—Sleep apnea is more frequent as age advances; sex. Apneas are more frequent among males in a ratio of approx. 3.5:1; neck circumference (in centimeters)—The neck circumference is an important predictive factor. The thicker the neck, the greater probability of apneas; BMI (Body Mass Index)—The BMI has a similar significance to that of the Neck Circumference, but is slightly less relevant; somnolence—Somnolence is a good indicator for sleep apnea; question response variables—The variables are coded with reference to the section: G = general sleep question; R = respiratory related question; S = somnolence related question. For all variables, high response values indicate greater probability of apneas. All variables, with the exception of R9, coded as: 1 “never”; 2 “rarely”; 3 “sometimes”; 4 “often”; 5 “always”. R9 is coded as: 1 “none”; 2 “once”; 3 “twice”; 4 “more than twice”.

The selected questionnaire questions and their descriptions are as follows: Question G3: Are you accustomed to taking a nap during the day? Question R1: Do you snore while asleep or have you been told that you do? Question R2: Does your snoring wake up your partner or can it be heard from another room? Question R6: Do you wake up at night with a sensation of choking? Question R7: Have you been told that you “stop breathing” when you are asleep? Question R8: Has your partner woken you for fear that you have stopped breathing? Question R9: How many times do you get up to go to the toilet at night? Question R10: Do you sweat a lot at night? Question R12: Do you wake up with a dry mouth? Question S4: Do you fall asleep in meetings or in public places? Question S5: Do you fall asleep while eating? Question S7: Do you fall asleep when driving on the motorway? Question S9: Do you fall asleep when driving if you stop at a traffic light?
Table 1
Basic statistics for the clinical variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Frequencies for categorical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>22</td>
<td>86</td>
<td>52.94</td>
<td>50 male, 21 female</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td>40 = no, 28 = yes, 3 = unknown</td>
</tr>
<tr>
<td>Neck circum. (cm)</td>
<td>34</td>
<td>50</td>
<td>39.52</td>
<td></td>
</tr>
<tr>
<td>BMI (body mass index)</td>
<td>19</td>
<td>43</td>
<td>25.46</td>
<td></td>
</tr>
<tr>
<td>Somnolence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDI index (output)</td>
<td>0</td>
<td>85</td>
<td>19</td>
<td>39 Positive cases; 32 Negative cases</td>
</tr>
</tbody>
</table>

**Question S10:** Do you fall asleep in your workplace while doing your normal work activities? (Refer to Table 1 for basic statistics of the clinical variables)

6. Results

6.1. Questionnaire responses—comparison of categorical and scalar representation of questions

In this section we evaluate the responses to the questions with the categorical form, and the responses to the questions with the scalar form. We compare the response frequencies to identify tendencies, differences, and improvements, if any, of the scalar form over the categorical form.

With reference to Table 2, we observe from the ‘Sca’ columns that in general, the fact that a person tends to think of a response in a scalar form rather than categorical, depends more on the question than the linguistic label (never, rarely, ...). Notwithstanding, if we study subgroups of questions (G, R, S) we can see signs of greater frequencies for the ‘Sca’ responses ‘frequently/always’ (R), and ‘never/rarely’(S). In Table 2 we can also see clear tendencies for specific questions, such as S9 with a higher frequency on responses ‘never’ and ‘rarely’, and R12 for the preference for higher range values ‘sometimes’, ‘frequently’ and ‘always’. We can also see an inversion of the tendency for responses to ‘never’ and ‘rarely’ when we compare categorical and scalar response frequencies (totals at bottom of respective columns).

6.2. Diagnosis using aggregation function—some initial results

We give the clinical questionnaire data and the meta-information to the aggregation operator which ‘fuses’ all its inputs together in one single diagnosis output per patient. We have tried four variations: (i) scalar question representation; (ii) categorical question representation; (iii) ‘relevance’ weight assignment by medical expert; (iv) ‘relevance’ weights learned by genetic algorithm. From the complete Apnea dataset of 71 cases, 41 were randomly sampled for the training set and 30 were randomly sampled for the test set. The resulting diagnostic accuracy of permutations of these techniques executed against the test set is given for positive, negative and all cases in Table 3.

With reference to Table 3, we observe a typical result with greater accuracy for positive cases and lesser accuracy for the negative cases, with the expert assigned weights giving slightly better results than the genetically learned ones. The results compare favorably
with the literature [13–15] for pure questionnaire based diagnosis of sleep apnea syndrome which tends to be in the order of 55–65% accuracy, and pure clinical data based diagnosis which is in the order of 70–90%. We think that, giving the genetic algorithm more evolutive time (we used only 15 generations) and a bigger population (we used 80 individuals) would give a better result for the learned weights.

6.3. Comparison of predictive accuracy of diagnosis using wowa aggregation with other predictive modelling methods

In order to compare the method with other artificial intelligence predictive techniques we executed a neural network and a tree induction algorithm against the same data, to predict the degree of apnea-hypopnea (RDI). The neural network was a standard feed-forward net with 3 layers, and the rule induction was run with unlimited tree depth and minimum of 5% of cases to form a branch. As in Section 6.2, we divided the data into a random sampled 58% training set (41 cases) and 42% test set (31 cases). The results of executing against the test set are summarized in Table 4, in which we see that WOWA aggregation performs better than neural nets and tree induction overall and for positive cases. For negative cases, WOWA performs worse than tree induction and slightly better than neural nets. In general neural nets and tree induction techniques require larger data volumes in order to build models, whereas the weighted aggregation approach should produce reasonable results with much fewer cases. We will have to look more closely at the reasons for

Table 2
Summary of frequencies categorical responses to each question (Cat) and the number of scalar questions responded as scalar (as opposed to a categorical response) (Sca)

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Frequently</th>
<th>Always</th>
<th>(M)issing</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cat</td>
<td>Sca</td>
<td>Cat</td>
<td>Sca</td>
<td>Cat</td>
<td>Sca</td>
<td>Cat</td>
</tr>
<tr>
<td>G3</td>
<td>13</td>
<td>8</td>
<td>16</td>
<td>14</td>
<td>20</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>R1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>R2</td>
<td>12</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>14</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>R6</td>
<td>39</td>
<td>9</td>
<td>4</td>
<td>15</td>
<td>18</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>R7</td>
<td>37</td>
<td>7</td>
<td>2</td>
<td>13</td>
<td>15</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>R8</td>
<td>42</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>R9</td>
<td>20</td>
<td>6</td>
<td>16</td>
<td>26</td>
<td>25</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>R10</td>
<td>11</td>
<td>6</td>
<td>21</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>R12</td>
<td>14</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>18</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>S4</td>
<td>49</td>
<td>11</td>
<td>3</td>
<td>12</td>
<td>11</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>S5</td>
<td>41</td>
<td>11</td>
<td>5</td>
<td>13</td>
<td>8</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>S7</td>
<td>61</td>
<td>13</td>
<td>6</td>
<td>15</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>S9</td>
<td>54</td>
<td>10</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>S10</td>
<td>48</td>
<td>7</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Totals</td>
<td>445</td>
<td>115</td>
<td>116</td>
<td>180</td>
<td>158</td>
<td>115</td>
<td>153</td>
</tr>
</tbody>
</table>

* Mainly omitted by people who indicated that they do not drive a car.
weaker performance in the negative cases of the apnea data, and we are in the process of collecting new cases to double the size of the train and test data sets.

7. Conclusions

From the questionnaire responses we can see interesting tendencies emerging of the way in which patients respond to the questions, depending on the type of the question, and the strength of the response required. In some cases a question can provoke more of a scalar response (‘shades of gray’) while in other cases the question provokes a more ‘black or white’ response from the patient. With respect to the diagnostic accuracy, we can see a promising result, achieved with few cases and a wide dimensionality of problem (19 variables). We have also been able to include three types of meta-data as part of the processing, thus adding insight which may improve the end result. As medical informaticians, we have learnt that careful selection of an adequate medical application is fundamental; one criteria for choosing an application is that it must allow real scope for improvement with respect to existing methods. Also, collaboration with medical experts has as prerequisite, the need for sufficient availability of their time for initial definition of the meta-data, selection of attributes and later analysis of the feedback of the results. The data quality and how representative a sample is, are also key aspects, together with the challenge of obtaining and capturing real case data in situ from the hospital environment. The authors are grateful to Dr Marina Rodriguez and the staff of the Sleep Center of the Hospital de la Santisima Trinidad, Salamanca (Spain), for their participation, and to Vicenc Torra of the Institute for Investigation in Artificial Intelligence, Bellaterra (Spain).

Table 3
Diagnostic accuracy on test dataset for positive, negative and all cases

<table>
<thead>
<tr>
<th></th>
<th>Positive cases</th>
<th>Negative cases</th>
<th>All cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical question representation/weights(^b) assigned by medical expert</td>
<td>0.735(^a)</td>
<td>0.462</td>
<td>0.498</td>
</tr>
<tr>
<td>Categorical question representation/weights(^b) learned by genetic algorithm</td>
<td>0.645</td>
<td>0.374</td>
<td>0.530</td>
</tr>
<tr>
<td>Scalar question representation/weights(^b) assigned by medical expert</td>
<td>0.625</td>
<td>0.433</td>
<td>0.598</td>
</tr>
<tr>
<td>Scalar question representation/weights(^b) learned by genetic algorithm</td>
<td>0.601</td>
<td>0.459</td>
<td>0.550</td>
</tr>
</tbody>
</table>

\(^a\) Correlation coefficients of predicted RDI with real RDI values.
\(^b\) Relevance weights.

Table 4
Comparison of the predictive accuracy of neural net, tree induction and WOWA algorithms with the test apnea dataset

<table>
<thead>
<tr>
<th></th>
<th>Neural net</th>
<th>Tree induction</th>
<th>WOWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cases (test)</td>
<td>0.540(^a)</td>
<td>0.548</td>
<td>0.598</td>
</tr>
<tr>
<td>Positive cases</td>
<td>0.600</td>
<td>0.523</td>
<td>0.735</td>
</tr>
<tr>
<td>Negative cases</td>
<td>0.450</td>
<td>0.625</td>
<td>0.462</td>
</tr>
</tbody>
</table>

\(^a\) Correlation coefficients of predicted RDI value and real RDI value.
8. Uncited references

[10,11]

References