Knowledge-Based Data Analysis: First Step Toward the Creation of Clinical Prediction Rules Using a New Typicality Measure

Mila Kwiatkowska, M. Stella Atkins, Najib T. Ayas, and C. Frank Ryan

***Abstract*—Clinical prediction rules play an important role in medical practice. They expedite diagnosis and limit unnecessary tests. However, the rule creation process is time consuming and expensive. With the current developments of efficient data mining algorithms and growing accessibility to medical data, the creation of clinical rules can be supported by automated rule induction from data. A data-driven method based on the reuse of previously collected medical records and clinical trial statistics is cost-effective; however, it requires well defined and intelligent methods for data analysis. This paper presents a new framework for knowledge representation for secondary data analysis and for generation of a new typicality measure, which integrates medical knowledge into statistical analysis. The framework is based on a semiotic approach for contextual knowledge and fuzzy logic for approximate knowledge. This semio-fuzzy framework has been applied to the analysis of predictors for the diagnosis of obstructive sleep apnea. This approach was tested on two clinical data sets. Medical knowledge was represented by a set of facts and fuzzy rules, and used to perform statistical analysis. Statistical methods provided several candidate outliers. Our new typicality measure identified those, which were medically significant, in the sense that the removal of those important outliers improved the descriptive model. This is a critical preprocessing step towards automated induction of predictive rules from data. These experimental results demonstrate that knowledge-based methods integrated with statistical approaches provide a practical framework to support the generation of clinical prediction rules.**

***Index Terms*—Clinical prediction rules (CPRs), data analysis, fuzzy logic, obstructive sleep apnea (OSA), typicality measure.**

I. Introduction

CLINICAL prediction rules (CPR) are used by medical practitioners as formal guidelines in diagnosis, prognosis, and treatment [1]. The rules simplify and expedite diagnosis and treatment for serious cases demanding immediate attention, and limit unnecessary diagnostic tests for low-probability cases. The rules provide quantitative predictive measures using factors from medical history, physical examination, and laboratory tests [2]. However, before the rules can be utilized in medical practice, they must be created, validated, and evaluated in clinical settings [3], [4]. By its nature, the creation of CPRs is time consuming and resource intensive. However, with the recent availability of electronic patient records and access to medical databases, the process of rule creation can be supported by machine learning methods providing automated or semiautomated rule induction from data [5]–[7].

Such a data-driven approach reuses existing data sets collected for medical research and clinical records of diagnosed patients. The secondary use of medical data reduces the cost of data acquisition, provides access to rare medical cases, and allows for analysis of diversified populations. On the other hand, secondary analysis of data from heterogeneous sources presents several challenges [8], [9].

- *Sampling bias:* Clinical studies use diverse collecting methods, inclusion criteria, and sampling methods.
- *Referral bias:* Most clinical studies are based on patients referred to specialists by the primary-care practitioners; therefore, the data represent a preselected group with a high prevalence of the disease.
- *Selection bias:* Clinical data sets include patients with different demographics such as gender, age, and ethnicity.
- *Method bias:* Clinical studies and patient records use diverse types and numbers of measurements and definitions of outcome. Therefore, predictors have varied specifications, granularities, and precisions.
- *Clinical spectrum bias:* Patient records represent varied severity of a disease and co-occurrence of other medical problems.

Therefore, the reuse of medical data requires intelligent data analysis as part of the preprocessing step in the knowledge discovery (KD) process [10]. This step must explicitly incorporate the medical knowledge required for the interpretation of the measurements and handling of the imprecision and uncertainty inevitably present in medical data.

In this paper, we present a new knowledge-based framework for the secondary analysis of heterogeneous data. Our *semio-fuzzy* framework is based on a semiotic approach for the contextual interpretation of the predictors and a fuzzy logic for the representation of the imprecision of measurements [11]. We applied the semio-fuzzy framework in analysis of the predictors used in the diagnosis of obstructive sleep apnea (OSA). As a first step towards the creation of prediction rules, we have constructed a prototype for a medical knowledge base (KB) and used it for data analysis from two dissimilar sources: a clinical research study and a database of patients' records.

Manuscript received December 22, 2005; revised July 25, 2006. This paper was supported in part by the New Investigator Award from CIHR/BC Lung Association, in part by the Scholar Award from the Michael Smith Foundation, and in part by the Departmental Scholar Award from the UBC.

M. Kwiatkowska is with the Computing Science Department, Thompson Rivers University, Kamloops, BC V2C 5N3, Canada (e-mail: mkwiatkowska@tru.ca).

M. S. Atkins is with the Computing Science Department, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: stella@cs.sfu.ca).

N. T. Ayas and C. F. Ryan are with the Faculty of Medicine, University of British Columbia, Vancouver, BC V5Z 3J5, Canada (e-mail: nayas@vanhosp.bc.ca; fryan@interchange.ubc.ca).

Digital Object Identifier 10.1109/TITB.2006.889693

We concentrated on four tasks: analysis of atypical values, analysis of monotonic patterns, analysis of statistical outliers, and comparison of atypical values and statistical outliers. The results show that traditional statistical data analysis is not sufficient for identification of atypical values in heterogeneous data sets obtained through various data collecting methods and nonrandom sampling techniques.

The paper is organized as follows. Section II provides a brief background information on OSA, clinical prediction rules, and predictors. Section III introduces the semiotic and fuzzy logic framework for predictor representation. We define the concept of typicality and describe a Fuzzy Inference System (FIS) for an anatomical typicality measure of an important OSA predictor, neck thickness. Section IV describes the clinical data sets. Section V presents experimental results including the analysis of typicality, monotonic patterns, and statistical outliers. Section VI provides the discussion and conclusion.

II. Background: Clinical Prediction Rules in the Diagnosis of OSA

A. OSA

OSA is a common and serious respiratory disorder afflicting approximately 2%–4% of the population [12]. OSA is caused by the repetitive collapse of the soft tissues in the throat as the result of the natural relaxation of muscles during sleep. The soft tissue blocks the air passage and the sleeping person literally stops breathing (apnea event) or experiences a partial obstruction (hypopnea event). Apnea occurs only during sleep and is, therefore, a condition that might go unnoticed for years. The gold standard for the diagnosis of OSA is an overnight in-laboratory polysomnography (PSG) study involving several recordings: electroencephalogram, electrocardiogram, electromyogram, airflow, breathing effort, and oxygen saturation.

OSA is associated with other medical conditions such as hypertension, congestive heart failure, and stroke [13]. Although the diagnosis of OSA using PSG is relatively straightforward and treatment is readily available, a large segment of the population is not diagnosed because of limited access to the overnight PSG. Therefore, patients suffering from OSA might spend several months waiting for the diagnosis. However, several clinical studies demonstrated that clinical prediction rules can be successfully used for initiation of an early OSA treatment for very severe cases (before formal diagnosis by PSG) [14], for preassessment of symptomatic patients by general practitioners (in this case, rules are often used in combination with overnight at-home oximetry) [15], and for prioritization of OSA cases for an urgent PSG [16]–[20].

B. Prediction Rules and Predictors

In general, the clinical prediction rules can be described as IF-THEN rules or arithmetic formulas for calculation of OSA probability based on particular predictors. A predictor, in a medical context, is defined as an established or suspected symptom, sign, correlate, or comorbid condition. The diagnostic predictors for OSA involve six factors:

1) anatomical signs such as obesity and large neck;
2) nocturnal symptoms of snoring and breathing pauses;
3) diurnal symptoms of excessive daytime sleepiness;
4) demographic factors such as gender and age;
5) coexisting medical conditions such as hypertension and diabetes; and
6) lifestyle factors such as smoking [21].

In this paper, we investigate an important OSA predictor, neck thickness, which is measured clinically as a neck circumference (NC) [22]. We study the association between NC and body mass index (BMI), calculated as weight (in kilograms) divided by the squared height (in meters).

III. Framework for Knowledge Representation

A. Semio-Fuzzy Approach

To represent the medical knowledge, we have used a framework combining *semiotics* and *fuzzy logic*, i.e., a *semio-fuzzy* approach. This framework addresses the imprecision of predictors and contextualization of the predictor interpretation in the diagnostic process.

A semiotic approach has been introduced into our model using Peirce's semiotic triangle to represent the concept, representation, and interpretation [23], [24]. Therefore, predictors are defined at three levels: conceptualization, representation (operationalization), and interpretation. The conceptualization level defines the medical concept (ontology) in terms of its general semantics. The representation level defines the possible measures of the medical concept. The interpretation level involves three aspects: a diagnostic value of the predictor, a practical utility of the predictor in a clinical setting, and a knowledge base.

The KB for the predictors has two components: the Fact Repository and the Typicality Measure System. The repository contains four types of facts: purpose (e.g., diagnostic, prognostic, treatment evaluation), context (e.g., specific subgroups), bias (e.g., dependencies between predictors), and view (diagnostic criteria used by the clinics).

B. Predictor Definition

We define the predictor as a quadruple, $\langle C, M, I, KB\rangle$. The predictor conceptualization is represented by C, the set of applicable measures for the concept by M, and the possible interpretations by I. A knowledge base for the predictor is represented by KB.

The interpretation I is defined as a pair, $I = \langle V, U\rangle$, where V represents the diagnostic value of the predictor and U represents the utility of the predictor. The diagnostic value comprises, for example, test specificity, sensitivity, and positive predictive value. The utility of the predictor involves pragmatic aspects such as test cost, difficulty, and health risks.

Fig. 1 shows the semiotic triangle of concept (neck thickness), representation (neck circumference), and interpretation defined by four aspects: purpose, context, bias, and view.

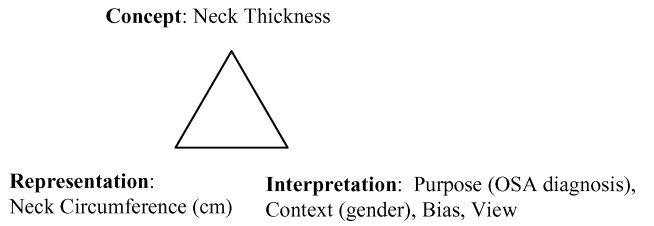

Fig. 1. Knowledge representation using Peirce's semiotic triangle.

C. Fact Repository for the Predictors

Based on published medical studies [17], [25]–[33], we created a fact repository for neck thickness. The repository contains seven facts (KB1–7) grouped into four aspects of interpretation:

1) *Purpose:* Diagnostic significance
   KB1—A large neck is a characteristic sign of OSA [17], [25]–[33].
2) *Context:* Subpopulation differences
   KB2 (gender)—In general, the NC is significantly larger in men than in women [17], [27]–[29], [31].
   KB3 (ethnicity)—In general, Asians tend to have smaller necks than whites [17].
3) *Bias:* Monotonic dependencies
   KB4—Heavier people are expected to have larger necks. However, the regional distribution of fat is gender-specific. BMI measures generalized obesity, while the larger neck reflects upper-body obesity, typical for men [27]. NC was shown to be smaller in women despite a larger BMI [31].
   KB5—Taller people are expected to have larger necks [27].
4) *View:* Measurement domain
   KB6—Neck circumference is measured at the level of the cricothyroid membrane, using a measuring tape [17].
   KB7—NC ranges from 25 to 65 cm for adults. Male NCs can be divided into three groups (the ranges were selected for the diagnosis of OSA): small to normal (<42), intermediate (42–45), and large (>45) [26].

D. Typicality Measure for the Predictors

The concept of *typicality* has been studied extensively for the classification of objects, especially graphical and image objects. Recent studies described a typicality measure for shapes and colors for image segmentation and retrieval [34], [35]. In general, typicality describes how well a given object fits into a specific category. The typicality measure depends on the representation of the category: rule-based, prototype-based, or exemplar-based [36], [37]. In the rule-based representation, a category is defined as a set of Boolean rules, which can be evaluated to true (object belongs to the category) or false (object is not in the category). In the prototype-based representation, the category is defined by the most dominant member (prototype) and the typicality is measured by the similarity of an object to the prototype. In the exemplar-based representation, the category is defined by all exemplars from a given category and the typicality is measured by the similarity to all exemplars.

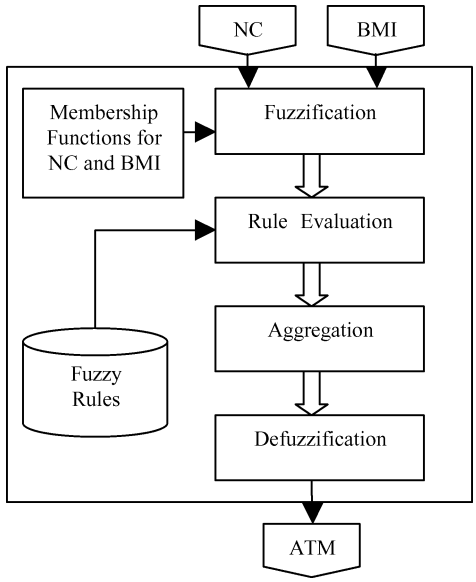

Fig. 2. FIS for ATM.

This paper proposes a semio-fuzzy typicality measure based on prior medical knowledge represented explicitly by fuzzy rules and membership functions. The typicality measure ranges from 0 to 1. Low values (e.g., 0.1) represent atypical values, whereas high values, (e.g., 0.9) represent typical values. This approach provides a practical method to detect medically atypical values among data from heterogeneous samples and a knowledge-based method to identify medically significant values among statistical outliers.

E. FIS for the Anatomical Typicality Measure

We applied our semio-fuzzy typicality measure to determine the anatomical typicality of neck circumference in relationship to BMI. We constructed a prototype of FIS using Fuzzy Logic Toolbox for MATLAB v. 7.01. Fig. 2 shows the FIS with two inputs: NC and BMI and one output, the anatomical typicality measure (ATM). Using the Mamdani inference process [38] the input values are fuzzified according to predefined membership functions (Figs. 3 and 4). Subsequently, fuzzy rules are evaluated and the results are aggregated into a single fuzzy set. Finally, the results are defuzzified (centroid) to produce a single ATM value.

We defined NC and BMI as two linguistic variables.

The predictor NC is defined as PNC

$$\text{PNC} = \langle \text{NC}, T(\text{NC}), [25,\ 65], M \rangle \tag{1}$$

where NC is the name of the variable, $T(\text{NC})$ is the set of terms for NC: {small, typical, large, atypical}, an interval [25, 65] is the domain for NC (in centimeter), and M is the set of membership functions defining the terms, $M\{\mu_{\text{small}}, \mu_{\text{typical}}, \mu_{\text{large}}, \mu_{\text{atypical}}\}$. The membership functions for NC in the general population (Fig. 3) were created using KB7 and KB2.

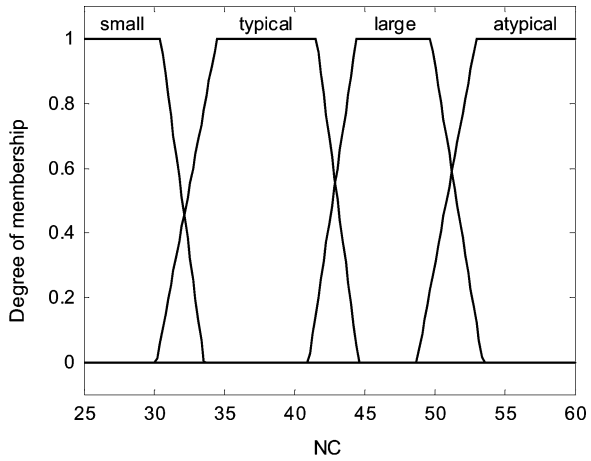

Fig. 3. Membership functions for four levels of NC in the general population.

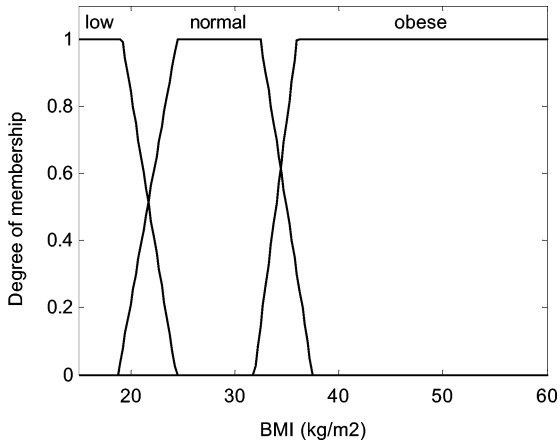

Fig. 4. Membership functions for three levels of BMI.

The predictor BMI is defined as PBMI

$$\text{PBMI} = \langle \text{BMI}, T(\text{BMI}), [15,\ 60], M \rangle \tag{2}$$

where BMI is the name of the variable, T(BMI) is the set of terms for BMI: {low, normal, obese}, an interval [15, 60] is the domain for BMI (in kilogram per square meter), and M is the set of membership functions, $M = \{\mu_{\text{low}}, \mu_{\text{normal}}, \mu_{\text{obese}}\}$. Fig. 4 shows the membership functions for the three levels of BMI. In general, the BMI > 30 is classified as obese. However, we adjusted the threshold value for the obesity to 35 kg/m^2 to accommodate the fact that most patients referred to sleep disorders clinic have a high BMI.

Based on the fact KB4, we created 12 fuzzy rules defining the anatomical typicality for NC in relationship to BMI. In general, an increased BMI should result in increased neck circumference and vice versa: a low BMI should correlate with smaller neck. For example, the following three rules define ATM for small NC and three levels of BMI:

Rule 1) If NC is *small* and BMI is *low*, then ATM is *high*.

Rule 2) If NC is *small* and BMI is *normal*, then ATM is *medium*.

Rule 3) If NC is *small* and BMI is *obese*, then ATM is *low*.

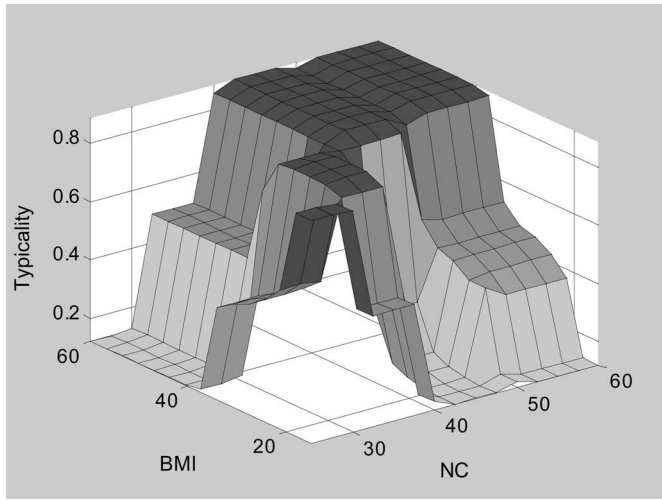

Fig. 5. Surface plot for anatomical typicality measure based on NC and BMI.

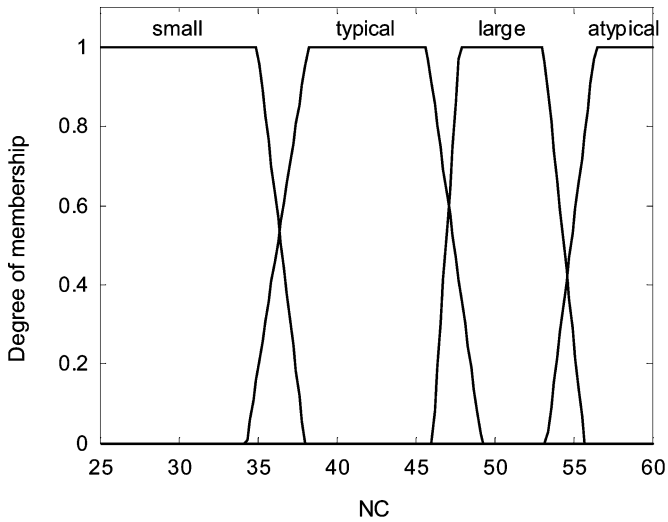

Fig. 6. Membership functions for four levels of NC adapted for males.

Fig. 5 shows a surface plot for ATM for the variable NC in relation to the variable BMI. The typicality is measured on a scale from 0 to 1.0. A low BMI and high NC significantly reduce the typicality measure, and a very high BMI and high NC elevate the typicality measure.

F. Contextual Interpretation of Typicality: Gender-Specific Membership Functions for NC

Based on knowledge fact KB2, which states that males have significantly larger necks than females, we created two additional sets of NC membership functions for males and females, as shown in Figs. 6 and 7, respectively. Using the general and gender-specific membership functions, the FIS calculates three ATM values: ATM$_{\text{G}}$ for general population, ATM$_{\text{M}}$ for males, and ATM$_{\text{F}}$ for females.

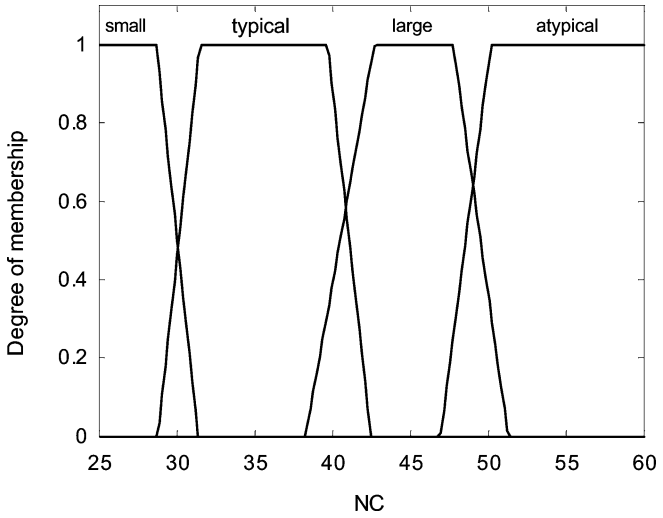

Fig. 7. Membership functions for four levels of NC adapted for females.

IV. Clinical Data Sets

The results presented in this paper are based on two data sets: A ($N = 239$) and B ($N = 147$). Data set A was collected for the clinical study of correlation between craniofacial measurements and OSA [17]. Data set B contains data from an educational sleep disorder clinic. These two data sets represent two categories of data mining sources: data collected for a clinical study and medical records of patients.

A. Data Set A

Data set A contains the data of 239 patients: 199 males (83%) and 40 females (17%); 164 Asian patients (69%) and 75 white patients (31%). The records have a mean age of 48.5 ($\pm$SD, $\pm$12.0) (middle aged) and mean BMI 29.2 ($\pm$5.7) (moderately obese).

B. Data Set B

Data set B contains the records of 147 consecutive patients: 104 males (71%) and 43 females (29%). The records have a mean age of 51.7 ($\pm$12.4) (middle aged) and a mean BMI of 33.9 ($\pm$7.0) (moderately obese to obese).

C. Comparison of Data Sets A and B

Table I shows the mean, standard deviation (SD), minimum, and maximum values for NC in data sets A and B. Data set A has lower mean and SD values than set B for all subgroups of patients.

In our study, we used three variables: NC, BMI, and gender. Additionally, we used the variable ethnicity available in set A.

V. Experimental Results

Knowledge-based data analysis concentrated on four tasks: analysis of atypical values, analysis of statistical outliers, comparison of statistical outliers and atypical values, and analysis of monotonic patterns.

TABLE I
Mean ($\pm Sd$), Minimum, and Maximum Values for Neck Circumference (cm) in Data Sets A and B

	Data set A	*Data set B*
All Patients NC	N = 239	N = 147
Mean ±SD (min; max)	39.9 ± 3.5 (31; 50)	42.9 ± 4.7 (32; 58)
Females NC	n = 40	n = 43
Mean ±SD (min; max)	36.3 ± 3.7 (31; 46)	39.1 ± 4.8 (32; 53)
Males NC	n = 199	n = 104
Mean ±SD (min; max)	40.7 ± 3.2 (31; 50)	44.4 ± 4.6 (39; 58)

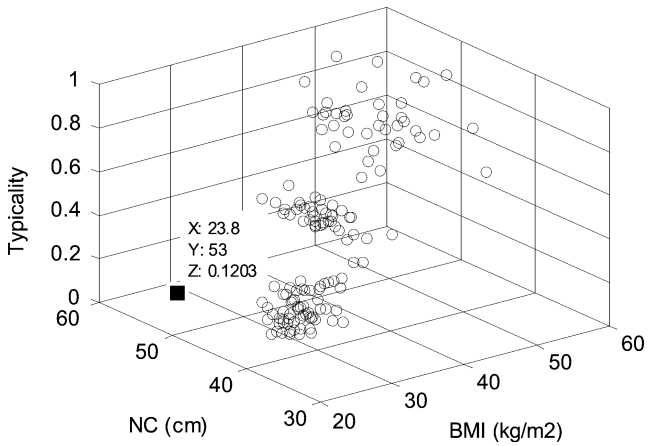

Fig. 8. Scatter plot for ATM based on NC and BMI (data set B). The *black square* represents a very low typicality, $Z = 0.1208$.

TABLE II
Typical NC Values (ATM ≤ 0.50) in Data Sets A and B

Data Set	*ID*	*Gender*	*BMI (kg/m²)*	*NC (cm)*	ATM_G	ATM_M	ATM_F
Set A	16	M	22.6	30.8	0.47	0.47	0.47
	239	M	34.5	50.0	0.50	0.12	0.50
	217	F	33.2	43.0	0.50	0.12	0.50
	32	F	40.8	39.0	0.50	0.50	0.50
Set B	101	F	23.8	53.0	0.12	0.12	0.12
	106	F	50.1	37.0	0.44	0.43	0.47

A. Analysis of Atypical Values

The NC values for data set A and B were analyzed for their anatomical typicality, in terms of their relationship to BMI. The FIS, described in Section III-E, calculated the ATM values for each record from data sets A and B.

Fig. 8 shows the general typicality $\mathrm{ATM_G}$ (Z-axis) based on BMI (X-axis) and NC (Y-axis). A data point marked by a black square represents an atypically large neck circumference ($Y = 53$ cm) for a low BMI ($X = 23.8$), with a low typicality value $Z = 0.1208$.

We identified six anatomically atypical NC values using a threshold of $\mathrm{ATM_G} \leq 0.50$. Table II shows the atypical records from data sets A and B. IDs are unique numbers within each data set.

The very low typicality value seen in Fig. 8 corresponds to the female patient ID 101.

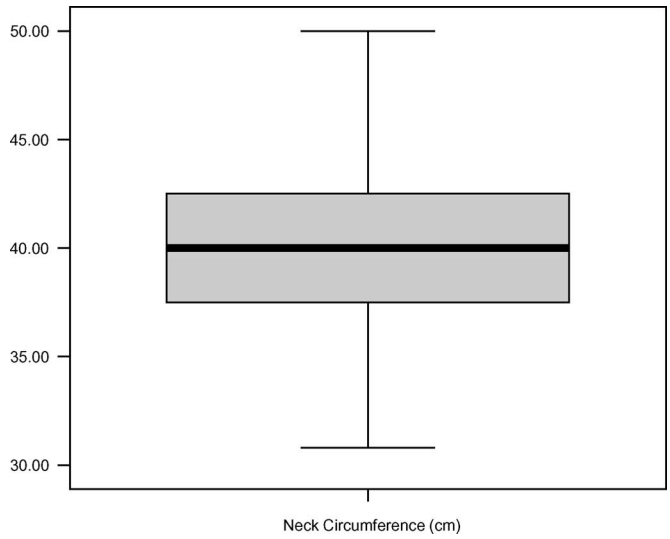

Fig. 9. Box plot for neck circumference (all patients, data set A).

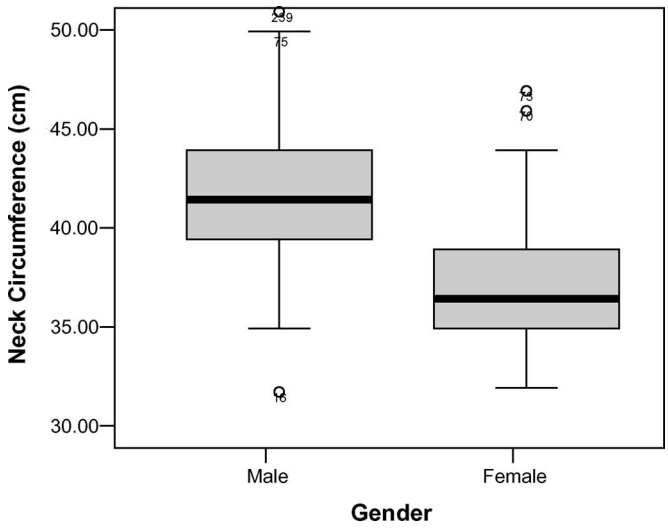

Fig. 10. Box plots for NC by gender (data set A).

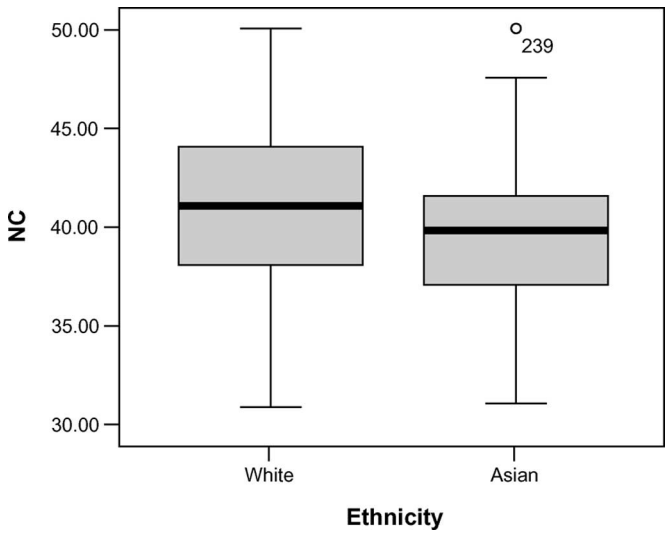

Fig. 11. Box plots for NC by ethnicity (data set A).

B. Analysis of Statistical Outliers

In general, outliers in a data set can be defined after Barnett and Lewis [39] as "an observation (or subsets of observations) which appears to be inconsistent with the reminder of that set of data." There are several causes for outliers: data entry errors, measurement errors, natural variation in studied population, or unusual values due to the sampling error.

Outlier detection is an important subject studied in statistics and, recently, in the context of data mining and machine learning [40]–[44]. The analysis of medical outliers involves three steps: identification, explanation, and handling. Outliers in medical data must be detected and individually evaluated since they may represent interesting anomalies or exceptional cases. The statistical methods have been successfully used for the detection of outliers in homogenous samples. However, medical data reuse involves data sets and samples from heterogeneous groups. Thus, we propose to combine purely statistical methods with a knowledge-driven approach.

In this study, we performed statistical outlier detection in two phases: univariate analysis and bivariate analysis (residual and distance-based). In this section, we present the results from the univariate outlier analysis and describe a knowledge-driven approach for creation of the subgroups. In Section V-C, we summarize our findings from the bivariate analysis.

We performed a univariate outlier analysis based on a median value, interquartile range (IQR), and box plot diagrams [45]. Outliers are shown as small circles and are identified by IDs—unique numbers within each data set.

1) Univariate Outlier Analysis for Data Set A: Fig. 9 shows the median and quartile values for NC for all patients in data set A. As seen in the figure, the box plot method did not identify outliers. Thus, we applied the facts stored in the KB. Based on KB2, we grouped the data by gender, as shown in Fig. 10.

Based on KB3, we grouped the data by ethnicity (Fig. 11). Based on facts KB2 and KB3, we grouped the data by gender and by ethnicity (Fig. 12). It is seen from Fig. 10 that males have much larger NC than females (confirming fact KB2) and that there is one male outlier with a low NC (ID 16) and two with high NCs (IDs 75 and 239). There are two female outliers, both with very large NCs (IDs 70 and 73). Fig. 11 shows only one outlier (ID 239) that also has been identified in subgroups based on gender.

Fig. 12 shows three outliers: two outliers detected by gender and ethnicity subgrouping (IDs 16 and 239), and one new outlier (ID 217) of an Asian female with a relatively large NC.

Division of data into gender and ethnic subgroups provided three male outliers (IDs 16, 75, 239) and three female outliers (IDs 70, 73, 217).

2) Univariate Outlier Analysis for Data Set B: Fig. 13 shows the median and quartile values for NC for all patients in set B. There are three outliers: IDs 137, 21, and 129.

Based on KB2, we grouped the data by gender. Fig. 14 shows three male outliers (indicated also by Fig. 13) and an additional female outlier with a very large NC (NC $= 53$), ID 101.

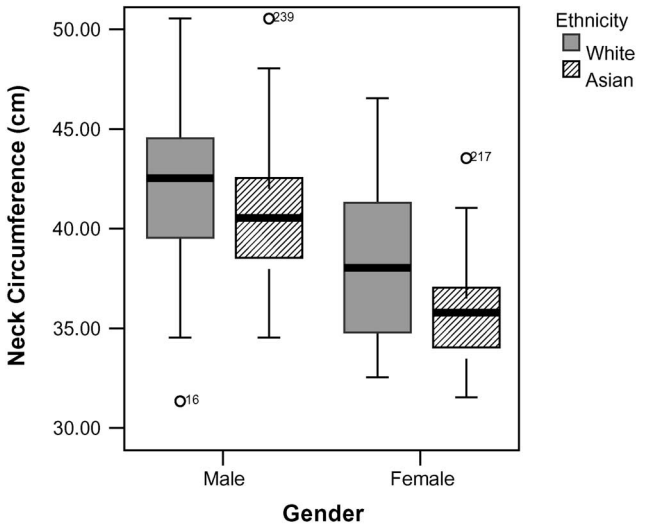

Fig. 12. Box plots for NC by gender and ethnicity (data set A).

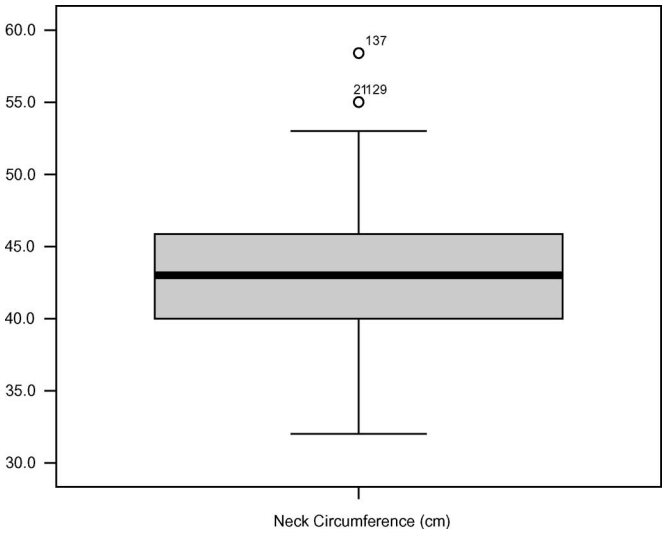

Fig. 13. Box plot for neck circumference (all patients, data set B).

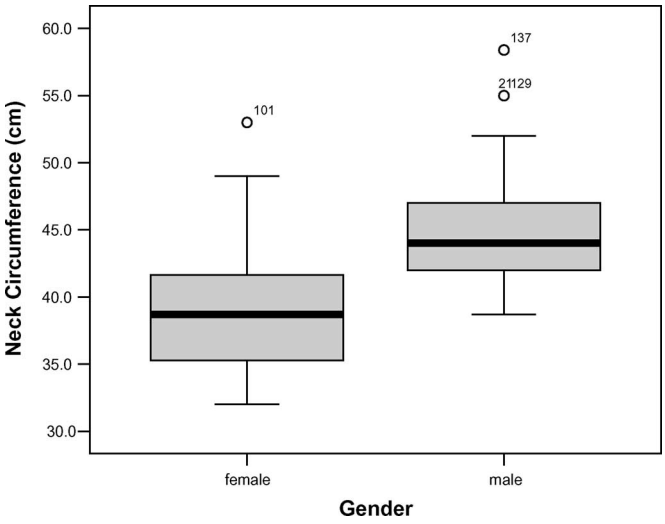

Fig. 14. Box plots for neck circumference by gender (data set B).

TABLE III
STATISTICAL OUTLIERS AND THEIR ATMS IN DATA SET A

Phase	*ID*	*Gender*	*BMI*	*NC*	ATM_G	ATM_M	ATM_F
Phase I	16	M	22.6	31	0.47	0.47	0.47
	75	M	44.1	50	0.89	0.88	0.88
	239	M	34.5	50	0.50	0.12	0.50
	70	F	49.8	45	0.83	0.88	0.87
	73	F	58.3	46	0.89	0.88	0.88
	217	F	33.2	43	0.50	0.12	0.50
Phase II*	32	F	40.8	39	0.50	0.42	0.50

* Only bivariate outliers with $ATM_G = 50$ are reported.

TABLE IV
STATISTICAL OUTLIERS AND THEIR ATMS IN DATA SET B

Phase	*ID*	*Gender*	*BMI*	*NC*	ATM_G	ATM_M	ATM_F
Phase I	21	M	52.9	55	0.89	0.88	0.88
	129	M	39.0	55	0.83	0.16	0.50
	137	M	51.4	58	0.83	0.88	0.88
	101	F	23.8	53	0.12	0.12	0.12
Phase II*	106	F	50.1	37	0.42	0.42	0.43

* Only bivariate outliers with $ATM_G = 50$ are reported.

C. Comparison of Statistical Outliers and Atypical Values

We studied medical outliers using two concepts: medically atypical values and statistical outliers. In Section V-A, we identified 6 atypical NC records with $ATM_G \leq 0.50$ in data sets A and B (Table II). In Phase I, described in Section V-B, we used a combination of a simple univariate analysis with a knowledge-driven grouping of data. As a result, we identified 10 univariate outliers in both sets. In Phase II, we performed the statistical outlier detection using bivariate analysis (residual analysis) and distance-based analysis (Mahalanobis distance). As a result, we identified 21 additional outliers: 9 by the residual analysis and 12 by the Mahalanobis distance. Then, we used the $ATM_G \leq 0.50$, to select medically significant bivariate outliers.

1) Typicality Measures for the Outliers in Data Set A: Table III shows the statistical outliers and their ATMs. Phase I identified six outliers and Phase II detected 1 additional outlier with $ATM_G \leq 0.50$, ID 32.

We applied the threshold value of $ATM_G \leq 0.50$ to outliers identified in Phase I and obtained three atypical outliers: IDs 16, 239, and 217. Thus, from both phases, we identified four anatomically atypical outliers in data set A: IDs 16, 32, 217, and 239.

2) Typicality Measures for the Outliers in Data Set B: Table IV shows the statistical outliers and their ATMs. Phase I identified four outliers: three with high ATMs and one with a low ATM. The IDs 21, 129, and 137 have relatively high anatomical typicality. Although $NC > 50$ cm is large, this value is not unusual among the obese to severely obese male patients. The record ID 101 of a slim ($BMI < 25$) female has low values for three ATMs. This indicates that a very large neck ($NC = 53$) is unusual for a slim patient. For Phase II, the table shows only one additional outlier with $ATM_G \leq 0.50$, ID 106.

We applied the threshold value of $ATM_G \leq 0.50$ to the outliers identified in Phase I and obtained only one outlier: ID 101.

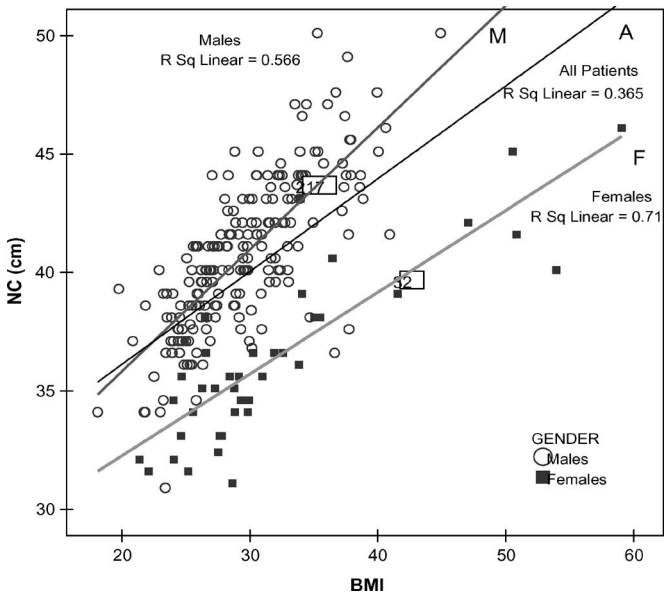

Fig. 15. Scatter plot of NC versus BMI for males and females in data set A.

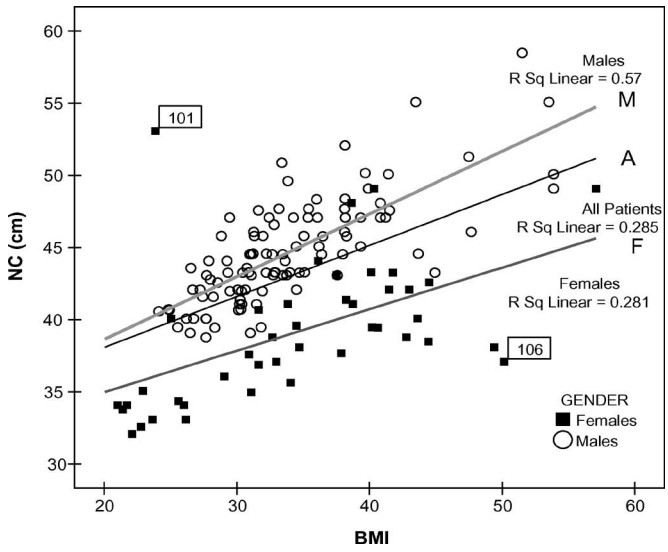

Fig. 16. Scatter plot of NC versus BMI for males and females in data set B.

Thus, from both phases, we identified two anatomically atypical outliers in data set B: IDs 101 and 106.

D. Analysis of Monotonic Patterns

The monotonic patterns for NC and BMI are based on the fact KB4: Heavier people are expected to have larger necks. We studied the relationship between BMI and NC using the Pearson's correlation coefficient r and the coefficient of determination r^2 [45]. The value of r^2 describes the proportion of the variation in the values of NC, which can be explained by BMI.

1) Correlation Analysis for Data Set A: Fig. 15 shows a scatter plot for NC versus BMI. Based on facts KB2 and KB4, we marked the data points by gender: males are indicated by open circles and females by black squares. We also added the least-squares (LS) regression lines for males (line M), females (line F), and all patients (line A). The anatomically atypical outliers ($\mathrm{ATM_G} \leq 0.50$) for females are marked by rectangles with IDs.

The values of r^2 (R Sq Linear in Fig. 15) are noticeably different for each of the studied groups. For all patients, the BMI explains only 37% of the variability of NC; however, the explanation power of BMI increases when the data are grouped by gender. For the male group, the BMI values explain 57% of NC variability. For the female group, the BMI values explain 71% of NC variability.

2) Correlation Analysis for Data Set B: Fig. 16 shows a scatter plot for NC versus BMI. Similar to Fig. 15, we used the circles and squares to indicate gender and added LS regression lines for males (M), females (F), and all patients (A). The value of r^2 is higher for the male group than for all patients. For males, the BMI explains 57% of the variability of NC. Surprisingly, the value of r^2 for females is lower than for all patients. For females, the BMI explains 28% of the variability of NC, whereas for all patients the BMI explains 28.5%.

TABLE V
COEFFICIENT OF DETERMINATION OF NC VERSUS BMI IN DATA SETS A, B, AND SUBGROUPS

	Data set A (N = 239) r^2	*Data set B (N = 147)* r^2	*Data sets A+B (N = 386)* r^2
All Patients	0.37 (239)	0.29 (147)	0.38 (386)
Males	0.57 (199)	0.57 (104)	0.63 (303)
Females	0.71 (40)	0.28 (43)	0.45 (83)

3) Comparison of the Correlation Between BMI and NC in Data Sets A and B: We compared the strength of the correlation between BMI and NC in both data sets (all correlations are significant, $p < 0.01$). Table V shows the r^2 for all patients and subgroups of patients. In data set A, the correlation between BMI and NC is stronger for both subgroups than for all patients. In data set B, the correlation is stronger for the male group than for all patients but weaker for the female group. The correlations in combined data set $(\mathrm{A} + \mathrm{B})$ exhibit the pattern: higher values for both subgroups and lower for all patients. Note that the correlation in the female group in set B is noticeably lower than in set A. However, the correlations for the male group are the same in both sets. Thus, our further analysis concentrates on the female groups.

4) Medical Outliers and their Effects on Correlation: We studied the correlation between BMI and NC for the female group in data sets A and B, and we examined the influence of four outliers with $\mathrm{ATM_G} \leq 0.50$ (Table II): IDs 217 and 32 from data set A (see Fig. 15) and IDs 101 and 106 from data set B (see Fig. 16). Table VI shows the r and r^2 values calculated with and without the outliers. In data set A, the removal of ID 217 increased only slightly the strength of the correlation, while the removal of ID 32 did not change the correlation. However, in data set B, the removal of ID 101 increased the strength of the correlation from $r^2 = 0.28$ to $r^2 = 0.50$, and

TABLE VI
CORRELATIONS OF NC VERSUS BMI FOR FEMALES IN DATA SETS A AND B

		NC versus BMI	
		r	r^2
Set A	Females (n = 40)	0.84**	0.71
	Without ID 217	0.87**	0.76
	ID 32	0.84**	0.71
	IDs 217 & 32	0.87**	0.76
Set B	Females (n = 43)	0.53**	0.28
	Without ID 101	0.71**	0.50
	ID106	0.57**	0.33
	ID 101 & 106	0.75**	0.57

** Correlation is significant at the 0.01 level (two-tailed).

the removal of both outliers, IDs 101 and 106, increased the strength to $r^2 = 0.57$. Interestingly, the value 0.57 is the same as the correlation strength for the male group in both data sets.

VI. DISCUSSION AND CONCLUSION

Clinical prediction rules are important and practical guidelines used in diagnosis, prognosis, and treatment. However, the creation process of the prediction rules is lengthy and expensive. In this paper, we have discussed a cost-effective method of reusing existing medical data from clinical studies and patient's records. Furthermore, we have demonstrated that the reuse and integration of data from heterogeneous data sources requires explicit representation of the predictors, their measures, and their interpretations. We have described a new framework based on semiotics and fuzzy logic for knowledge representation and secondary data analysis.

In this paper, we have concentrated on four important issues in the preprocessing step of the KD process: 1) identifying the medically atypical values; 2) finding the statistical outliers using a combination of traditional statistical methods with a knowledge-driven approach; 3) evaluating the medical typicality of statistical outliers; and 4) determining the relationships between predictors.

We addressed the first issue of medical typicality by introduction of an anatomical typicality measure based on FIS. We calculated ATM values for NC from two data sets A and B, and selected six anatomically atypical records based on $\mathrm{ATM_G} \leq 0.50$.

The second issue was analyzed based on the medical facts stored in the KB created by us for the neck-thickness predictor. We observed that the statistical outliers are sensitive to the medical variations of the subgroups of patients, e.g., the outliers identified among female patients (e.g., 101 in set B) are not outliers in the general population.

The third issue was addressed by evaluating the anatomical typicality of identified outliers. We used three methods for outlier analysis: univariate, bivariate (residual analysis), and distance-based (Mahanalobis). We identified 10 univariate and, additionally, 21 bivariate outliers in both data sets. We classified six statistical outliers as medically atypical by applying the threshold of $\mathrm{ATM_G} \leq 0.50$. Interestingly, the six atypical outliers are the same as the six anatomically atypical records identified by FIS. We observed that the identification of outliers within a single data set is sensitive to the particular distribution of the data in that set. For example, univariate outlier ID 239 from data set A would not be detected by univariate analysis in data set B. This outlier would be detected by residual analysis but only after grouping by gender. However, the typicality measure provides a consistent quantification of typicality based on prior medical knowledge. Thus, the outlier ID 239 would have the same ATM values in any set of data.

The fourth issue was examined by analysis of monotonic patterns in context of different subgroups of patients and removal of medically atypical outliers. We have shown that the coefficient of determination r^2 significantly changes for the three groups: the entire population, males, and females. Furthermore, the removal of an outlier with a low ATM value (e.g., ID 101 in data set B) noticeably increases the r^2 value (from 0.28 to 0.50).

Our results demonstrate that the proposed semio-fuzzy framework can be successfully utilized for secondary analysis of medical data. Moreover, this framework provides a uniform approach to predictor representation.

We would like to apply these methods to generate clinical prediction rules for apnea diagnosis and evaluate these rules for sensitivity and specificity on other clinical data sets. Furthermore, we would like to use these techniques in other medical domains, including other respiratory disorders such as asthma, and in nonmedical domains such as psychology and cognitive science.

REFERENCES

[1] M. H. Ebell, *Evidence-Based Diagnosis: A Handbook of Clinical Prediction Rules*. New York: Springer, 2001.
[2] A. Laupacis, N. Sekar, and I. G. Stiell, "Clinical prediction rules. A review and suggested modifications of methodological standards," *JAMA*, vol. 277, no. 6, pp. 488–494, Feb. 1997.
[3] I. G. Stiell, G. H. Greenberg, G. A. Wells, I. McDowell, A. A. Cwinn, N. A. Smith, T. F. Cacciotti, and M. L. A. Sivilotti, "Prospective validation of a decision rule for the use of radiography in acute knee injures," *JAMA*, vol. 275, no. 8, pp. 611–615, Feb. 1996.
[4] T. McGinn, G. Guyatt, P. Wyer, D. Naylor, I. G. Stiell, and S. Richardson, "Users' guides to the medical literature XXII: How to use articles about clinical decision rules," *JAMA*, vol. 284, no. 1, pp. 79–84, 2000.
[5] W. Duch, R. Adamczak, K. Grabczewski, G. Zal, and Y. Hayashi, "Fuzzy and crisp logical rule extraction methods in application to medical data," in *Computational Intelligence and Applications*, P. Szczepaniak, Ed. Berlin, Germany: Springer-Verlag, 2000, pp. 593–616.
[6] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[7] A. Kusiak, J. A. Kern, K. H. Kernstine, and B. T. L. Tseng, "Autonomous decision-making: A data mining approach," *IEEE Trans. Inform. Technol. Biomed.*, vol. 4, no. 4, pp. 274–284, Dec. 2000.
[8] D. K. Owens and H. C. Sox, "Medical decision-making: Probabilistic medical reasoning," in *Medical Informatics: Computer Applications in Health Care and Biomedicine*, 2nd ed., E. H. Shortliffe and L.E. Perreault, Eds. New York: Springer, 2001, pp. 76–129.
[9] M. J. Druzdzel and F. J. Diez, "Combining knowledge from different sources in causal probabilistic models," *J. Mac. Learn. Res.*, vol. 4, pp. 295–316, 2003.
[10] K. J. Cios and G. W. Moore, "Medical data mining and knowledge discovery: Overview of key issues," in *Medical Data Mining and Knowledge Discovery*, K. J. Cios, Ed. Heidelberg, Germany: Springer-Verlag, 2001, pp. 1–20.
[11] M. Kwiatkowska and M. S. Atkins, "Information fusion in the diagnosis of obstructive sleep apnea: A semio-fuzzy approach," in *Proc. NAFIPS Conf.*, 2004, pp. 55–60.

[12] N. J. Douglas, *Clinicians' Guide to Sleep Medicine*. London, U.K.: Arnold, 2002.
[13] C. R. F. Nieto, T. Young, B. K. Lind, E. Shahar, J. Samet, S. Redline, R. B. D'agostino, A. B. Newman, M. D. Lebowitz, and T. G. Pickering, "Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study," *JAMA*, vol. 283, no. 14, pp. 1829–1836, 2000.
[14] N. Pelletier-Fleury, N. Meslier, F. Gagnadoux, C. Person, D. Rakotonanahary, H. Ouksel, B. Fleury, and J.-L. Racineux, "Economic arguments for the immediate management of moderate-to-severe obstructive sleep apnoea syndrome," *Eur. Respir. J.*, vol. 23, pp. 53–60, 2004.
[15] N. Roche, B. Herer, C. Roig, and G. Huchon, "Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea," *Chest*, vol. 121, no. 3, pp. 747–752, 2002.
[16] P. J. Ryan, M. F. Hilton, D. A. Boldy, A. Evans, S. Bradbury, S. Sapiano, K. Prowse, and R. M. Cayton, "Validation of British Thoracic Society guidelines for the diagnosis of the sleep apnoea/hypopnea syndrome: Can polysomnography be avoided?," *Thorax*, vol. 50, pp. 972–975, 1995.
[17] B. Lam, M. S. M. Ip, and C. F. Ryan, "Craniofacial profile in Asian and white subjects with obstructive sleep apnoea," *Thorax*, vol. 60, pp. 504–510, 2005.
[18] R. Rosenberg and S. A. Mickelson, "Obstructive sleep apnea: Evaluation by history and polysomnography," in *Snoring and Obstructive Sleep Apnea*, 3rd ed., D. N. F. Fairbanks, S. A. Mickelson, and B. T. Woodson, Eds. Philadelphia, PA: Lippincott Williams & Wilkins, 2003, pp. 39–50.
[19] N. J. Douglas, "Home diagnosis of the obstructive sleep apnoea/hypopnoea syndrome," *Sleep Med. Rev.*, vol. 7, no. 1, pp. 53–59, 2003.
[20] I. Gurubhagavatula, G. Maislin, and A. I. Pack, "An algorithm to stratify sleep apnea risk in a sleep disorders clinic population," *Am. J. Respir. Crit. Care Med.*, vol. 164, pp. 1904–1909, 2001.
[21] T. Young, J. Skatrud, and P. E. Peppard, "Risk factors for obstructive sleep apnea," *JAMA*, vol. 291, no. 16, pp. 2013–2016, 2004.
[22] W. H. Tsai, J. E. Remmers, R. Brant, W. W. Flemons, J. Davies, and C. Macarthur, "A decision rule for diagnostic testing in obstructive sleep apnea," *Am. J. Respir. Crit. Care Med*, vol. 167, no. 10, pp. 1427–32, May 2003.
[23] T. A. Sebeok, *Signs: An Introduction to Semiotics*. Toronto, Canada: Univ. Toronto Press, 1999.
[24] P. B. Andersen, *A Theory of Computer Semiotics*. Cambridge: Cambridge Univ. Press, 1997, p. 11.
[25] R. P. Millman, C. C. Carlisle, S. T. McGarvey, S. E. Eveloff, and P. D. Levinson, "Body fat distribution and sleep apnea severity in women," *Chest*, vol. 107, no. 2, pp. 362–6, Feb. 1995.
[26] K. A. Ferguson, T. Ono, A. A. Lowe, C. F. Ryan, and J. A. Fleetham, "The relationship between obesity and craniofacial structure in obstructive sleep apnea," *Chest*, vol. 108, no. 2, pp. 375–81, Aug. 1995.
[27] D. R. Dancey, P. J. Hanly, C. Soong, B. Lee, J. Shepard, and V. Hoffstein, "Gender differences in sleep apnea: The role of neck circumference," *Chest*, vol. 123, pp. 1544–1550, 2003.
[28] A. T. Whittle, I. Marshall, I. Mortimore, P. K. Wraith, R. J. Sellar, and N. J. Douglas, "Neck soft tissue and fat distribution comparison between normal men and women by magnetic resonance imaging," *Thorax*, vol. 54, pp. 323–328, 1999.
[29] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: A population health perspective," *Am. J. Respir. Crit. Care Med.*, vol. 165, no. 9, pp. 1217–1239, 2002.
[30] C. F. Ryan and L. L. Love, "Mechanical properties of the velopharynx in obese patients with obstructive sleep apnea," *Am. J. Respir. Crit. Care Med.*, vol. 154, no. 3, pp. 806–812, Sep. 1996.
[31] J. A. Rowley, L. S. Aboussouan, and M. S. Badr, "The use of clinical prediction formulas in the evaluation of obstructive sleep apnea," *Sleep*, vol. 23, no. 7, pp. 929–938, 2000.
[32] J. A Rowley, C. S. Sanders, B. R. Zahn, and M. S. Badr, "Gender differences in upper airway compliance during NREM sleep: Role of neck circumference," *J. Appl. Physiol.*, vol. 92, pp. 2535–2541, 2002.
[33] A. S. Jordan and R. D. McEvoy, "Gender differences in sleep apnea: epidemiology, clinical presentation and pathogenic mechanisms," *Sleep Med. Rev.*, vol. 7, no. 5, pp. 377–389, 2003.
[34] S. Joshi, S. Pizer, P. T. Fletcher, P. Yushkevich, A. Thall, and J. S. Marron, "Multiscale deformable model segmentation and statistical shape analysis using medical descriptions," *IEEE Trans. Med. Imag*, vol. 21, no. 5, pp. 538–550, May 2002.
[35] S. Pizer, P. T. Fletcher, S. Joshi, A. Thall, J. Z. Chen, Y. Fridman, D. S. Fritsch, A. G. Gash, J. M. Glotzer, M. R. Jiroutek, C. Lu, K. E. Muller, G. Tracton, P. Yushkevich, and E. L. Chaney, "Deformable M-Reps for 3D medical image segmentation," *Int. J. Comput. Vis.*, vol. 55, no. 2–3, pp. 85–106, 2003.
[36] R. M. Nosofsky, "Exemplars, prototypes, and similarity rules," in *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes A. Healy*, vol. 1, S. Kosslyn and R. Shiffrin, Eds. Hillsdale, NJ: Erlbaum, 1992.
[37] J. P. Minda and J. D. Smith, "Prototypes in category learning: The effects of category size, category structure and stimulus complexity," *J. Exp. Psychol. Learn., Mem. Cogn.*, vol. 27, pp. 755–799, 2001.
[38] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*. Amsterdam, The Netherlands: Morgan Kaufmann, 2005.
[39] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 2nd ed. Norwich, U.K.: Wiley, 1987.
[40] X. Liu, G. Cheng, and J. X Wu, "Analyzing outliers cautiously," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 2, pp. 432–437, Mar.–Apr. 2002.
[41] V. Podgorelec, M. Hericko, and I. Rozman, "Improving mining of medical data by outliers prediction," in *Proc. Comput.-Based Med. Syst.*, Jun. 2005, pp. 91–96.
[42] M. K. Edwin, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *VLDB J.*, vol. 8, no. 3–4, pp. 237–253, 2000.
[43] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.
[44] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems," *Program. Comput. Softw.*, vol. 29, no. 4, pp. 228–237, Jul. 2003.
[45] B. H. Munro, *Statistical Methods for Health Care Research*. Philadelphia, PA: Lippincott Williams & Wilkins, 2001, p. 235.

Mila Kwiatkowska, photogaph and biography not available at the time of publication.

M. Stella Atkins, photogaph and biography not available at the time of publication.

Najib T. Ayas, photogaph and biography not available at the time of publication.

C. Frank Ryan, photogaph and biography not available at the time of publication.