Evaluation of clinical prediction rules using a convergence of knowledge-driven and data-driven methods: a semio-fuzzy approach

M. Kwiatkowska[1], N. T. Ayas[2,3] & F. Ryan[2]
[1]*Department of Computing Science, University College of the Cariboo, Canada*
[2]*Sleep Disorders Program, Faculty of Medicine, University of British Columbia, Canada*
[3]*Centre for Clinical Epidemiology and Evaluation, VCHRI, Canada*

Abstract

Researchers and practitioners in medicine use various clinical prediction rules to estimate the probability and severity of a disease. Based on a limited number of factors from medical history, physical examination, and laboratory tests, the practitioners use these rules to expedite diagnosis and treatment for serious cases and limit unnecessary tests for low-probability cases. However, before the rules can be used in clinics, they must be validated on large and diversified populations and evaluated in clinical settings. A computer system providing intelligent data analysis and data mining techniques can facilitate this lengthy and costly process. This paper describes a conceptual framework and a prototype system for the rule evaluation process. The study concentrates on predictive rules used in the diagnosis of obstructive sleep apnea (OSA), a common and serious respiratory disorder. The prototype system, Hypnos, includes (1) a framework for rule definition, and (2) a mechanism for rule evaluation. The rule definition framework is based on a semiotic and a fuzzy logic approach. The semiotic description incorporates a rule's syntax, semantics, and pragmatics. The fuzzy logic models the imprecise features. The rule evaluation mechanism supports first the validation and then comparison of the rules built by the domain experts and the rules generated by the data-driven method. The results from both methods are compared based on rule accuracy, interpretability, generality, and clinical utility. The prediction rules for OSA are described and evaluated using four datasets (1,300 records) from two clinics. The results show that rules obtained from data mining can confirm, contradict, or expand the rules created by medical experts. Therefore, the paper suggests a combination of knowledge-driven and data-driven methods for rule derivation and validation.
Keywords: clinical prediction rules, machine learning, decision trees, rule induction, obstructive sleep apnea, fuzzy logic, semiotics, diagnosis.



1 Introduction

Researchers and practitioners in medicine use various clinical prediction rules in diagnosis, prognosis, and treatment to estimate the probability and severity of a disease. Clinical prediction rules are based on a limited number of factors from medical history, physical examination, and laboratory tests. The rules are used to expedite the diagnostic process and initiate early treatment for severe cases or to avoid unnecessary testing in low probability cases. The development process for the predictive rules involves derivation, validation, and evaluation in clinical settings. The derivation requires successive refinements and tests using statistical techniques. Preliminary results from this study demonstrate that the computer system can support the derivation and validation of the rules. This paper describes a semio-fuzzy framework for the definition of predictive rules. It then presents a prototype based on this framework that supports two-way rule generation: (1) from hypotheses to data and (2) from data to hypotheses. The former, leading from human generated hypotheses to data sets, is based on the clinical experience of medical experts. The second approach is based on machine learning techniques, generating hypotheses from the data sets. The machine generated rules are interpreted and compared with the human generated hypotheses. This interactive process has an exploratory and confirmatory purpose: it allows for the discovery of new patterns from data and provides confirmation or contradiction of hypothetical rules.

This paper focuses on the application of clinical prediction rules in the diagnosis of obstructive sleep apnea (OSA). Section 2 provides a brief introduction to OSA and its diagnostic criteria. Section 3 discusses the semio-fuzzy framework. Section 4 describes the data. Section 5 presents the preliminary results from the prototype system, Hypnos. The last section provides the conclusions and the directions for future work.

2 Understanding of the problem domain

The background knowledge presented in this section is necessary to understand the mechanism of obstructive sleep apnea, the diagnostic criteria, and the significance of the clinical prediction rules in the diagnosis of OSA.

2.1 Obstructive sleep apnea

Obstructive sleep apnea is a common, serious respiratory disorder afflicting approximately 2-4% of the population. OSA is caused by 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 continuous recordings of EEG, ECG, EOG, EMG, airflow, breathing effort, and oxygen saturation. OSA



is associated with hypertension, congestive heart failure, stroke, and coronary artery disease. Furthermore, sleep deprivation caused by OSA leads to several social consequences, such as motor vehicle collisions, job related injures, and a decreased quality of life. 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 time factors, costs, and limited access to the overnight in-clinic PSG. Therefore, patients suffering from OSA might spend several months waiting for diagnosis. However, we believe that by using a combination of predictive rules and home studies, early treatment can be initiated in appropriate patients before formal diagnosis by PSG.

2.2 Diagnosis of OSA

The diagnosis of OSA uses two approaches: (1) a score of apnea/hypopnea events and (2) a combination of scoring and symptoms. Both approaches use an apnea-hypopnea index (AHI), calculated as a number of apnea and hypopnea events per hour of sleep [1]. An *apnea* is defined as a complete cessation of airflow for at least 10 seconds. A *hypopnea* is defined using various criteria consisting of one or more of the following three factors: partial reduction of airflow, oxyhemoglobin desaturation, and brief arousals from sleep. In the diagnosis based solely on the AHI index, apnea is classified as *mild* for AHI between 5 and 14.9, *moderate* for AHI between 15 and 29.9, and *severe* for AHI ≥ 30. However, the use of diverse scoring criteria for AHI calculation can result in significant differences in apnea diagnoses, especially for patients with low AHI scores [2,3]. Furthermore, the difficulty with the scoring of AHI is compounded by (1) natural night-to-night variations and (2) differences in diagnostic equipment. The results from the diagnostic process are documented in the form of a numeric outcome or a descriptive written report.

3 A semio-fuzzy framework for knowledge representation

A knowledge representation framework, based on a semiotic and fuzzy logic approach [4] defines two essential diagnostic concepts: *prediction rules* and *predictors*. The semiotic description defines the prediction rules at syntactic, semantic, and pragmatic levels. The fuzzy logic is used to model the inherent imprecision of the predictors using numerical, categorical, and linguistic values.

3.1 Prediction rules

The clinical prediction rule (CPR) is specified by an IF-THEN statement, a certainty factor, and usability. Hence, we define CPR as a triplet: $< RS, CF, U >$. Where the rule statement, *RS*, represents the rule's syntax, the rule certainty factor, *CF*, is a part of the rule's semantics, and the usability, *U*, determines the rule's pragmatic value.

3.1.1 Rule syntax

The rule is comprised of two parts: a *premise* and a *consequent*. The premise of the rule uses predefined predictors, for example, *age*, *gender*, or *hypertension*. A

proposition is a logical expression composed of a predictor variable, the relational operator (<, ≤, >, ≥, =), and a value; for example, *age* > *65*, *hypertension* = *yes*. The rules are in the conjunctive propositional form, for example, *age* > *65 AND gender* = *female*. The conclusion of the rule includes the class label. The rule statement is defined in extended BNF grammar, as follows:

<Rule statement> ::= IF <Rule premise> THEN <Rule consequent>
<Rule premise> ::= <Relational expression> {AND <Relational expression>}
<Relational expression> ::= <Predictor variable> <Relational operator> <Value>
<Relational operator> ::= < | ≥ | > | ≤ | =
<Value> ::= numerical value | categorical value | linguistic value
<Rule consequent> ::= class label

3.1.2 Rule semantics

The clinical prediction rule is a hypothetical statement with two functions: *descriptive* and *predictive*. In the *descriptive* sense, rules characterize the subpopulations of patients with higher or lower risks for the disease. In the *predictive* sense, rules assess the probability of a new patient belonging to one of the classes. The hypothetical quality of the rule is defined by the certainty factor (CF), a degree of belief ranging from -1.0 (absolute disbelief) to +1.0 (absolute belief), assigned to the rule by medical experts based on their clinical experience.

3.1.3 Rule pragmatics

The rule's pragmatic value is determined by three criteria: *internal validity*, *external validity*, and *clinical usability*. Internal validity is based on specificity and sensitivity. The external validity is based on rule generality: transferability to a different data set. The clinical usability comprises interpretability, simplicity, and practicality. The rule interpretability and practicality are qualitatively determined by the medical experts. The rule simplicity is measured, for example, by the length of the rule.

3.2 Predictors

A predictor is an established or suspected symptom, sign, correlate, or co-morbid condition. The Hypnos knowledge base stores 14 predictors divided into six categories: (1) anatomical signs: obesity, large neck size, and craniofacial and upper airway abnormalities, (2) nocturnal symptoms: snoring, breathing pauses, and choking, (3) diurnal symptoms: excessive daytime sleepiness, (4) demographic factors: gender, age, and familial aggregation, (5) coexisting medical conditions: hypertension and coronary artery disease; and (6) lifestyle factors: smoking and alcohol use [5,6,7]. The four predictors that apply to this study are *age*, *gender*, *hypertension* (HTN), and *obesity* measured by *body mass index* (BMI) in kg/m^2.

3.2.1 The fuzzy logic framework for predictors

Fuzzy logic allows for granulation of values and fuzzification of sharp borders between normal and abnormal values. For example, the hypertension predictor

can be represented by (1) categorical binary values: {*yes*, *no*} or (2) continuous numerical values for systolic (SBP) and diastolic blood pressure (DBP). In the case of categorical values, hypertension is defined as BP ≥ 140/90 mmHg, or current treatment with antihypertensive medications. However, this dichotomous division does not consider patients with high normal BP = 139/40, nor does it make a distinction between high and severely high blood pressures. In the case of continuous BP measure, the values can be divided into sharply defined intervals, for example, SBP < 140 mmHg, 140 ≤ SBP < 179 mmHg, and SBP ≥ 180 mmHg. The fuzzy logic approach defines five fuzzy sets of SBP: *low/normal*, *high normal*, *high*, and *severe*. Each set has a corresponding membership function, represented in figure 1. Thus, for example, SBP = 180 is partially *high* (degree of membership = 0.82) and partially *severe* (degree of membership = 0.82).

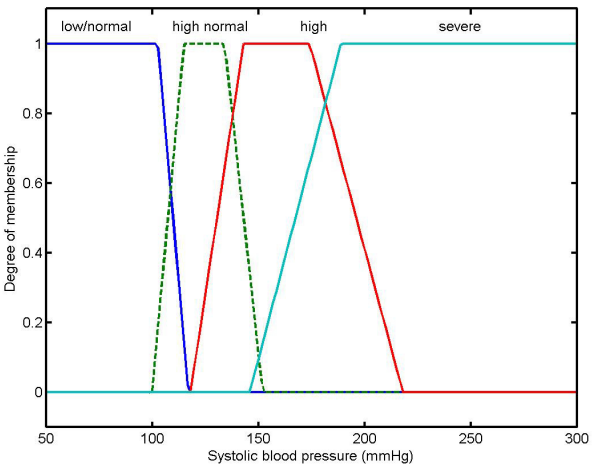


Figure 1: Trapezoid membership functions for systolic blood pressure (SBP).

4 Understanding the data

This study uses three data sets based on patients' records and one data set based on a study of healthy population. Data sets A (N=795) and B (N=298) were supplied by the Respiratory Clinic at Vancouver General Hospital; data sets C (N=196) and D (N=69) were provided by the Sleep Disorders Clinic at the University College of the Cariboo. Data sets A and B are based on standard PSG and use a uniform definition of AHI. However, data sets C and D involve diverse diagnostic techniques. Therefore, data sets A and B are used as the primary training and testing sets for the rule generation, set C is used for the text mining, and set D is used for the auxiliary analysis.

Set A has 795 records, males=539, females=256, mean age=50.4 years (STD=12.4), mean BMI=31.9 (STD=7.6), and hypertension=yes for 241 records.

Set B has 290 records, males=210, females=80, mean age=49.2 (STD=12.1), mean BMI=31.2 (STD=6.6), and hypertension=yes for 55 records.

The prevalence of OSA in sets A and B depends strongly on the AHI cut-off values. The changes in prevalence are illustrated by table 1. In our study, we use AHI ≥ 15 to define OSA, since this value typically indicates clinically important OSA requiring treatment. The records with AHI < 15 are classified as non-OSA.

Table 1: Changes in OSA prevalence for AHI cut-off 5, 10, and 15.

	Prevalence of OSA based on AHI cut-off values		
	AHI ≥ 5	AHI ≥ 10	AHI ≥ 15
Set A (N=795)			
OSA = yes	91.95% (731)	78.62% (625)	65.79% (523)
OSA = no	8.05% (64)	21.38% (170)	34.21% (272)
Set B (N=290)			
OSA = yes	83.10% (241)	62.41% (181)	47.93% (139)
OSA = no	16.90% (49)	37.59% (109)	52.07% (151)

Set C has 196 records, males=136, females=60, mean age=49.5 (STD=14.3), mean BMI = 30.81 (STD= 6.5), and hypertension=yes for 54 records. Set C is based on written narrative diagnoses, which were manually classified by a health practitioner as OSA (154), non-OSA (40), and inconclusive results (2). Each document is defined by a vector of 40 features grouped into three classes: (1) OSA synonyms, (2) confirmation class, and (3) disconfirmation class.

Set D contains 69 records of healthy physically fit adults, males=66 and females=3, mean age=41.8 (STD=10.0), mean BMI=27.88 (STD=3.1). The data regarding hypertension were inconsistent and were therefore excluded.

5 Preliminary results

The two-way rule generation involves (1) the knowledge-driven method, based on the hypothetical rules created by medical experts, (2) the data-driven method, based on machine-generated rules, and (3) the convergence of both methods. The preliminary results demonstrate that the rules obtained from data mining can confirm, contradict, or expand the rules created by medical experts.

This study applies two hypothetical rules from the knowledge-driven method, *ER1* and *ER2*, which exemplify (1) a high-risk group: older male patients with morbid obesity (BMI > 40), and (2) a low-risk group: young female patients with normal weight (BMI < 25):
ER1 = IF BMI>40 AND age>65 AND gender=male THEN OSA=yes (CF=0.9),
ER2= IF BMI<25 AND age<25 AND gender=female THEN OSA=no (CF=0.8).
Rule *ER1* has overall accuracy of 87.5% on set A and 89.3% on set B. Rule *ER2* has overall accuracy of 100% on set A and 50% on set B.

The data driven method is based on machine learning techniques generating classification rules. Since the clinical rules must be interpretable by medical experts, we are using techniques providing understandable rules: (1) decision tree generation and (2) rule induction.

The prototype system, Hypnos, uses three algorithms provided by the Weka repository: the decision tree learner J48 (C4.5), logistic model tree (LMT), and



the propositional rule induction algorithm JRIP [8]. The classifiers are trained and tested on set A using the stratified 10-fold cross-validation. This validation method divides data randomly into 10 stratified parts. Nine-tenths of the data is used for training and one-tenth for testing. This process is repeated 10 times and the overall error is calculated from the average of ten executions [8].

The written diagnoses from data set D are classified using text mining techniques. The classifier is generated using the decision tree learner J48 and tested by 10-fold cross-validation. The overall accuracy of the model is 90%; however, this accuracy might be too optimistic, since all diagnostic reports were prepared by the same sleep specialist. Therefore, this model requires further testing on diversified diagnostic texts.

5.1 Data mining

The data mining process involves three steps: (1) evaluation and ranking of the attributes, (2) generation of the classifiers, and (3) comparison of performance.

Evaluation and ranking of the attributes is performed using two methods: (1) chi-square and (2) information gain with respect to binary outcome (*OSA*, *non-OSA*). Table 2 represents the results for set A. The attribute evaluation is tested using 10-fold stratified cross-validation. Both methods produce the same ranking of the attributes: (1) *BMI*, (2) *hypertension* (HTN), (3) *age*, and (4) *gender*.

Table 2: Attributes evaluation using Chi-square and information gain.

	Chi-square		Information gain	
Attribute	Average merit	Average rank	Average merit	Average rank
BMI	21.727 +- 3.678	1.1 +- 0.3	0.022 +- 0.003	1.0 +- 0.00
HTN	10.065 +- 2.088	2.9 +- 0.83	0.010 +- 0.002	2.8 +- 0.87
Age	8.402 +- 8.414	2.9 +- 1.14	0.008 +- 0.008	3.1 +- 0.94
Gender	8.817 +- 2.183	3.1 +- 0.54	0.009 +- 0.002	3.1 +- 0.54

5.1.1 Generated classifiers

The machine-generated classifiers involve three predictive models: (1) Model 1, based on a decision tree, (2) Model 2, based on a logistic model tree, and (3) Model 3, based on the decision rules.

5.3.1.1 Model 1 Figure 2 represents the decision tree that was generated from the data in set A. The nodes represent the predictors, the branches correspond to predictor values, and the leaves correspond to the outcome classes. The two numbers in the leaves correspond to (1) instances covered by the rule premise and (2) exceptions from the rule.

5.3.1.2 Model 2 The model based on the Logistic Model Tree (LMT) [9] has two logistic regression functions: F_1 for non-OSA class and F_2 for OSA class:
$F_1 = -1.35 + age * 0.02 + gender * -0.53 + htn * 0.35 + bmi * 0.04$,
$F_2 = 1.35 + age * -0.02 + gender * 0.53 + htn * -0.35 + bmi * -0.04$.

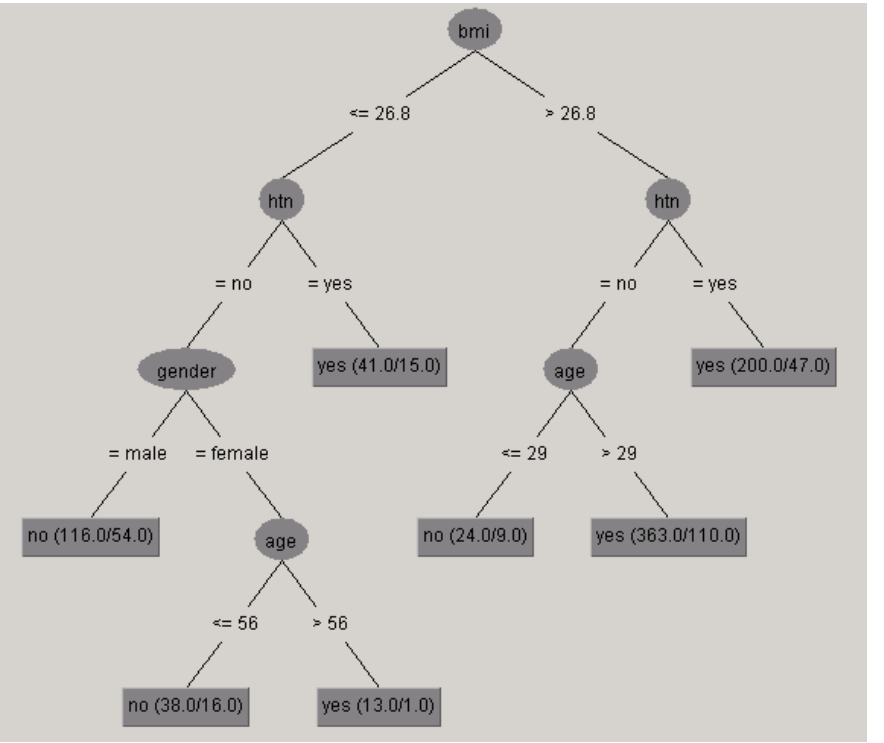


Figure 2: Decision tree trained and tested on data set A.

The instance *i* belongs to the *non-OSA* class with probability of $p_1(i)$, eqn (1).

$$p_1(i) = e^{F_1(i)} / (e^{F_1(i)} + e^{F_2(i)}). \quad (1)$$

5.3.1.3 Model 3 The model based on the propositional rule learner algorithm has three rules:
Rule 1: *IF (BMI ≤ 26.8) and (HTN = no) THEN OSA = no*,
Rule 2: *IF (HTN = yes) THEN OSA = yes*,
Rule 3: *IF (BMI > 26) THEN OSA= yes*.
Rule 3 has limited applicability for a general population. In set D, containing records of healthy, physically fit adults, 47 instances have BMI > 26.

5.2 Convergence of knowledge-driven and data-driven methods

The convergence of expert-generated models and machine-generated models is based on three criteria: (1) the equivalency of the predictors, (2) internal and external validity, and (3) clinical usability: interpretability, simplicity, and practicality. All models use the set of four predictors: *BMI*, *hypertension*, *age*, and *gender*. The accuracies of the machine-generated rules are shown in Table 3.

Table 3: Comparison of three machine-generated models.

	Model 1 DT	Model 2 LMT	Model 3 JRIP
Correctly Classified Instances	498	517	512
Incorrectly Classified Instances	297	278	283
Sensitivity	0.803	0.914	0.845
Specificity	0.287	0.143	0.257
Overall Accuracy	62.65%	65.03%	64.40%

The interpretation of computer-generated rules might (1) confirm the hypothetical rules, (2) provide contradictory examples, or (3) identify new insights. The following examples illustrate the three outcomes:

Example 1: Expert rule *ER1* specifying high OSA risks for morbidly obese older male patients:
ER1: IF BMI > 40 and age > 65 AND gender = male THEN OSA = yes
is confirmed by two rules obtained from Model 1:
R1: IF BMI > 26.8 AND HTN = yes THEN OSA=yes;
R2: IF BMI > 26.8 AND HTN= no AND age > 29 THEN OSA = yes.

Example 2: Expert rule *ER2* specifying low OSA risks for young female patients with normal weight:
ER2: IF BMI < 25 AND age < 25 AND gender = female THEN OSA = no
is contradicted by the rule generated from Model 1, which additionally includes *hypertension* (HTN): *IF BMI<=26.8 AND HTN=yes THEN OSA=yes*.

Example 3: A new insight is provided by Model 1, which divides female patients with normal BP and normal BMI (or slightly overweight) into two age groups: age <= 56 and age > 56. This specific age-based division could be associated with an increased risk of OSA among postmenopausal women.

6 Conclusions and future work

For some time medical researchers have been studying ways to improve the validity and reliability of clinical prediction rules in order to increase their accuracy, provide earlier treatment for OSA sufferers, and reduce the costs incurred by unnecessary testing. For this study we applied data mining techniques to test their ability to facilitate and refine the rule derivation and validation process in a new and innovative way. The convergence of the results from two-way rule generation confirms, contradicts, or expands the expert-generated prediction rules. Although limited only to four predictors, the results from this study demonstrate that our approach is valid, and warrants future work involving additional predictors and further data mining techniques. Our prototype system, Hypnos, successfully applied semio-fuzzy representation for prediction rules and complex predictors. Furthermore, it generated three models providing comprehensible rules.

In the course of this study, we identified two problems: (1) description of diverse definitions of OSA diagnostic criteria, and (2) the use of the data sets with high prevalence of OSA. The first problem was addressed in this study by

restricting the definition of OSA to a specific AHI threshold value (AHI≥15). The second problem must be addressed in future studies.

We are planning to expand our work in three directions: (1) development of models based on all 14 OSA predictors, (2) application of other machine learning techniques, for example, associative rules induction, (3) training and testing on larger and more diversified data sets. Furthermore, we are developing a telemedicine application, which will test the rules in a clinical setting.

References

[1] Douglas, N. J., *Clinicians' Guide to Sleep Medicine,* London: Arnold, 2002.

[2] Redline, S., Kapur, V. K., Sanders, M. H., Quan, S. F., Gottlieb, D. J., Rapoport, D. D., et al. , Effects of Varying Approaches for Identifying Respiratory Disturbances on Sleep Apnea Assessment. *Am. J. Respir. Crit. Care Med*., 161, pp. 369-374, 2000.

[3] Tsai, W. H., Flemons, W. W., Whitelaw, W. A. & Remmers, J. E., A Comparison of Apnea-Hypopnea Indices Derived from Different Definitions of Hypopnea. *Am. J. Respir. Crit. Care Med*., 159(1), pp. 43-48, 1999.

[4] Kwiatkowska, M. & Atkins, M. S., A Semio-Fuzzy Approach to Information Fusion in the Diagnosis of Obstructive Sleep Apnea. *Proc. of the 23rd International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2004), Banff*, pp. 55-60, 2004.

[5] Lamberg, L., Familial Link Seen in Obstructive Sleep Apnea. *JAMA*, 290(22), pp. 2925-2926, 2003.

[6] Nieto, C. R. F., Young, T., Lind, B. K., Shahar, E., Samet, J., Redline, S., et al., Association of Sleep-Disordered Breathing, Sleep Apnea, and Hypertension in a Large Community-Based Study. *JAMA*, 283(14), pp. 1829-1836, 2000.

[7] Young, T., Skatrud, J. & Peppard, P. E., Risk Factors for Obstructive Sleep Apnea. *JAMA*, 291(16), pp. 2013-2016, 2004.

[8] Witten, I. H. & Frank, E., *Data Mining: Practical Machine Learning Tools and Technologies with Java Implementations,* Morgan Kaufmann, 2000.

[9] Landwehr, N., Hall, M. & Eibe, F., *Logistic Model Trees*, Online. http://www.cs.waikato.ac.nz/~ml/publications/2003/landwehr-etal.pdf, 2003.

