*EXCLI Journal 2013;12:569-583 – ISSN 1611-2156*
*Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013*

## Original article:

# QUANTITATIVE POPULATION-HEALTH RELATIONSHIP (QPHR) FOR ASSESSING METABOLIC SYNDROME

Apilak Worachartcheewan[1,2], Chanin Nantasenamat[1,2]*,
Chartchalerm Isarankura-Na-Ayudhya[2],Virapong Prachayasittikul[2]*

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
* Corresponding authors:
E-mail: chanin.nan@mahidol.ac.th; phone: +66 2 441 4371 ext. 2720, Fax: +66 2 441 4380
E-mail: virapong.pra@mahidol.ac.th; phone: +66 2 441 4376, Fax: +66 2 441 4380

## ABSTRACT

***Background:*** Metabolic syndrome (MS) is a condition that predisposes individuals to the development of cardiovascular diseases and type 2 diabetes mellitus.
***Methods:*** A cross-sectional investigation of 15,365 participants residing in metropolitan Bangkok who had received an annual health checkup in 2007 was used in this study. Individuals were classified as MS or non-MS according to the International Diabetes Federation criteria using BMI cutoff of ≥ 25 kg/m$^2$ plus two or more MS components. This study explores the utility of quantitative population-health relationship (QPHR) for predicting MS status as well as discovers variables that frequently occur together. The former was achieved by decision tree (DT) analysis, artificial neural network (ANN), support vector machine (SVM) and principal component analysis (PCA) while the latter was obtained by association analysis (AA).
***Results:*** DT outperformed both ANN and SVM in MS classification as deduced from its accuracy value of 99 % as compared to accuracies of 98 % and 91 % for ANN and SVM, respectively. Furthermore, PCA was able to effectively classify individuals as MS and non-MS as observed from the scores plot. Moreover, AA was employed to analyze individuals with MS in order to elucidate pertinent rule from MS components that occur frequently together, which included TG+BP, BP+FPG and TG+FPG where TG, BP and FPG corresponds to triglyceride, blood pressure and fasting plasma glucose, respectively.
***Conclusion:*** QPHR was demonstrated to be useful in predicting the MS status of individuals from an urban Thai population. Rules obtained from AA analysis provided general guidelines (i.e. co-occurrences of TG, BP and FPG) that may be used in the prevention of MS in at risk individuals.

**Keywords:** metabolic syndrome, cardiovascular diseases, diabetes, data mining, QPHR

## INTRODUCTION

Metabolic syndrome (MS) is defined as a group of metabolic abnormalities comprising of central obesity, dyslipidemia, hyperglycemia, and hypertension (Babu and Fogelfeld, 2006). MS is associated with increased risks for cardiovascular diseases (CVD) (WHO, 2007) and type 2 diabetes mellitus (DM) (WHO, 2008). Therefore, the identification of MS is important for prevention of metabolism-associated diseases. Criteria used in the identification of

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

MS differed across organizations. Groups such as the World Health Organization (WHO, 1999), the European Group for the Study of Insulin Resistance (EGIR) (Balkau and Charles, 1999) in 1999, the National Cholesterol Education Program (NCEP) Adult Treatment Panel (ATP) III (NCEP, 2001) in 2001 and the International Diabetes Federation Central obesity (IDF) (Alberti et al., 2009) in 2005 have each established independent criteria for defining MS. Therefore, rapid identification of MS from the population would be beneficial for the prevention of CVD and type 2 DM.

Data mining is a robust tool for extracting useful knowledge from large quantities of data and can be readily applied to clinical data as to help physicians in the decision-making process of diagnosis, prognosis, and treatment of patients. Data mining techniques such as artificial neural network (ANN), support vector machine (SVM), multiple linear regression (MLR), principal component analysis (PCA), self organizing map (SOM), decision tree (DT) and association analysis (AA) have been successfully used in clinical medicine for predictive modeling of diseases (Chang et al., 2011; Firouzi et al., 2007; Kim et al., 2012a, b; Lee et al., 2000; Nahar et al., 2011; Obenshain, 2004; Ting et al., 2009; Worachartcheewan et al., 2010a; Yeh et al., 2011; Yoo et al., 2012). Furthermore, data mining have extensively been shown to be important tools in life sciences as it helps elucidate quantitative structure-activity relationships (QSAR) and quantitative structure-property relationships (QSPR) as a function of calculated physicochemical descriptors (Nantasenamat et al., 2005, 2007a, b, 2009, 2010; Prachayasittikul et al., 2010; Thippakorn et al., 2009; Worachartcheewan et al., 2011, 2012, 2013). The aim of this study is to apply DT, ANN, SVM, and PCA approaches in the development of quantitative population-health relationship (QPHR) models. Such predictive models were used in assessing health parameters of individuals from a large data set of urban Thai population as to predict their MS status as well as discover important variables contributing to MS by means of association rules.

## MATERIAL AND METHODS

### *Sample population*

A cross-sectional data set comprising of 15,365 individuals receiving an annual health check-up in 2007 from the Faculty of Medical Technology, Mahidol University in Bangkok, Thailand was previously reported by Worachartcheewan et al. (2010b). Such data set is comprised of anthropometric parameters along with blood pressures (measured according to standard procedures) and blood chemistry as analyzed at the Center of Medical Laboratory Services, Faculty of Medical Technology, Mahidol University. Individuals were categorized as MS according to IDF criteria (Alberti et al., 2009) using a cutoff of BMI $\geq$ 25 kg/m$^2$ (5,638 from total population) as the first component along with two or more components:

(1) blood pressure (BP) $\geq$ 130/85 mmHg or previously diagnosed hypertension,

(2) fasting plasma glucose (FPG) $\geq$ 100 mg/dL or previously diagnosed type 2 DM,

(3) triglyceride (TG) $\geq$ 150 mg/dL or specific treatment for triglyceride abnormality as well as high-density lipoprotein cholesterol (HDL-C) < 40 mg/dL in males or < 50 mg/dL in females or specific treatment for HDL-C abnormality.

Individuals with BMI $\geq$ 25 kg/m$^2$ were selected for QPHR study (encompassing a total of 5,638 individuals) as they met the first requirement of the IDF criteria of central obesity. From this subset of data, individuals with 2 or more MS components were identified as MS (2,991 individuals: 1,598 males and 1,393 females) while healthy individuals were classified as non-MS (2,647 individuals: 1,063 males and 1,584 females).

### *Metabolic abnormalities of MS*

Determining factors of MS (i.e. BP, FPG, TG and HDL-C) were stratified according to guidelines of the WHO (Wilson,

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

2009) and the IDF criteria (Alberti et al., 2009) as shown in Table 1. Furthermore, individuals having BMI ≥ 25 kg/m$^2$ were further divided into six BMI groups as well as separated by gender (male and female) and stratified into four age groups as presented in Table 1. These health parameters were used as input variables while the MS status (i.e. MS or non-MS) was used as the output variable.

**Table 1:** A stratification of the clinical and biochemical features in the urban Thai population

| Factors | | | MS (N=2,991) | | non-MS (N=2,647) | |
|---|---|---|---|---|---|---|
| | | | M | F | M | F |
| Gender | M | Male | 1,598 (53.43) | - | 1,063 (40.16) | - |
| | F | Female | - | 1,393 (46.57) | - | 1,584 (59.84) |
| Age (years) | Age1 | 20 - 34 | 67 (2.24) | 18 (0.60) | 83 (3.14) | 59 (2.23) |
| | Age2 | 35 - 44 | 410 (13.71) | 298 (9.96) | 381 (14.39) | 450 (17.00) |
| | Age3 | 45 - 54 | 739 (24.71) | 673 (22.50) | 419 (15.83) | 795 (30.03) |
| | Age4 | ≥ 55 | 382 (12.77) | 404 (13.51) | 180 (6.80) | 280 (10.58) |
| BMI | BMI1 | 25 - 25.9 | 362 (12.10) | 248 (8.29) | 348 (13.15) | 442 (16.70) |
| | BMI2 | 26 - 26.9 | 326 (10.90) | 236 (7.89) | 266 (10.05) | 347 (13.11) |
| | BMI3 | 27 - 27.9 | 245 (8.19) | 203 (6.79) | 170 (6.42) | 228 (8.61) |
| | BMI4 | 28 - 28.9 | 216 (7.22) | 165 (5.52) | 91 (3.44) | 184 (6.95) |
| | BMI5 | 29 - 29.9 | 146 (4.88) | 135 (4.51) | 73 (2.76) | 139 (5.25) |
| | BMI6 | ≥ 30 | 303 (10.13) | 406 (13.57) | 115 (4.34) | 244 (9.22) |
| BP | BP1 | < 130 and < 85 | 289 (9.66) | 279 (9.33) | 648 (24.48) | 1,070 (40.42) |
| | BP2 | 130 - 139 or 85 - 89 | 614 (20.53) | 549 (18.36) | 234 (8.84) | 280 (10.58) |
| | BP3 | 140 - 149 or 90 - 94 | 512 (17.12) | 434 (14.51) | 142 (5.36) | 186 (7.03) |
| | BP4 | 150 - 159 or 95 - 99 | 97 (3.24) | 50 (1.67) | 21 (0.79) | 25 (0.94) |
| | BP5 | ≥ 160 and ≥ 100 | 86 (2.88) | 81 (2.71) | 18 (0.68) | 23 (0.87) |
| FPG | FPG1 | <100 | 589 (19.69) | 543 (17.85) | 930 (35.13) | 1,415 (53.46) |
| | FPG2 | 100 - 125 | 783 (26.18) | 682 (22.80) | 120 (4.53) | 152 (5.74) |
| | FPG3 | ≥ 126 | 226 (7.56) | 177 (5.92) | 13 (0.49) | 17 (0.64) |
| TG | TG1 | < 150 | 362 (12.10) | 550 (18.39) | 865 (32.68) | 1,449 (54.74) |
| | TG2 | 150 - 199 | 518 (17.32) | 482 (16.12) | 92 (3.48) | 89 (3.36) |
| | TG3 | 200 - 299 | 547 (18.29) | 307 (10.26) | 84 (3.17) | 43 (1.62) |
| | TG4 | 300 - 399 | 166 (5.55) | 53 (1.77) | 22 (0.83) | 3 (0.11) |
| | TG5 | ≥ 400 | 5 (0.17) | 1 (0.03) | 0 (0) | 0 (0) |
| HDL-CM | HDL-CM1 | ≥ 40 | 1,171 (39.15) | - | 1,047 (39.55) | - |
| | HDL-CM2 | < 40 | 427 (14.28) | - | 16 (0.61) | - |
| HDL-CF | HDL-CF1 | ≥ 50 | - | 690 (23.07) | - | 1,463 (55.27) |
| | HDL-CF2 | < 50 | - | 703 (23.50) | - | 121 (4.57) |

Numbers in parentheses represent values in percentage. MS: metabolic syndrome, non-MS: non-metabolic syndrome, BMI: body mass index (kg/m$^2$), BP: blood pressure (mmHg), FPG: fasting plasma glucose (mg/dL), TG: triglyceride (mg/dL), HDL-CM and HDL-CF: high-density lipoprotein cholesterol (mg/dL) in male and female, respectively.

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

***Data pre-processing***

Independent variables were adjusted to comparable scale by standardizing variables to zero mean and unit variance. Standardization of variables was performed as described by the following equation:

$$x_{ij}^{stn} = \frac{x_{ij} - \bar{x}_j}{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 / N} \quad (1)$$

where $x_{ij}^{stn}$ is the standardized value, $x_{ij}$ is the value of each sample, $\bar{x}_j$ is the mean of each variables, and $N$ is the sample size of the data set.

***Quantitative population-health relationship (QPHR) modeling***

Health parameters from annual health check-ups of an urban Thai population served as the data set for multivariate analysis where individuals were classified as MS or non-MS by means of several data mining techniques.

***Decision tree analysis***

Decision tree (DT) is a supervised technique for classifying data into categorical classes of interest and the wisdom gained from the learning process are summarized in the form of if-then rules. DT finds the most important independent variable and sets it as the root node, which is followed by a series of bifurcating nodes when decision criteria are met. This is performed iteratively until leaf or terminal nodes are reached where it is then assigned one of many possible class labels of the dependent variable (i.e. MS or non-MS). This study employs the J48 algorithm (Witten et al., 2011), which is WEKA's implementation of the C4.5 DT learning algorithm. A confidence factor of 25 % was implemented and used in this study.

***Artificial neural network***

Artificial neural network (ANN) is a data mining technique that functions in a similar manner to the learning process of neurons in the human brain. ANN is essentially comprised of 3 layers of nodes: input, hidden and output layers (Zupan and Gasteiger, 1999). ANN parameters (i.e. number of hidden layer, learning epochs, learning rate and momentum) were optimized in an empirical manner as to obtain an optimal set of values. The back-propagation implementation (Nantasenamat et al., 2007b) of WEKA, version 3.4.5 (Witten et al., 2011), was employed in this study.

***Support vector machine***

Support vector machine (SVM) is a statistical learning method developed by Vapnik and co-workers (Cortes and Vapnik, 1995; Vapnik, 1998).This study employs John Platt's Sequential Minimal Optimization of the WEKA software package for SVM classification (Witten et al., 2011). It is essentially based on the principles of Structural Risk Minimization, which is a non-parametric and supervised classifier employing kernel functions for generating the transformation space. The radial basis function (RBF) kernel was employed in this study. Parameter optimization was performed by investigating the following two parameters: the $C$ and $\gamma$ parameters. This was performed in a two-step process that entails an initial course grid search followed by a more refined local grid search of optimal regions deduced from the coarse grid search (Worachartcheewan et al., 2011; Nantasenamat et al. 2013).The essence of SVM involves the mapping of data onto a high-dimensional feature space by means of kernel transformation in the form of $K(x,x_i)$ (Cortes and Vapnik, 1995; Vapnik, 1998). Samples were then classified into two separate classes by constructing hyperplanes that linearly separates the data. The optimal separating hyperplane for classifying the data is the hyperplane that maximizes the margins as to achieve maximum distance between the plane and nearest data. The classification process performs optimization of the Lagrange multipliers $\alpha_i$ with constraints $0 \leq \alpha_i \leq C$ and $\sum \alpha_i \gamma_i = 0$ in obtaining the decision function:

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

$$f(x) = sign(\sum_{i=l}^{l} \alpha_i y_i K(x, x_i) + b) \quad (2)$$

where $y_i$ represents input class labels (having values of -1 or 1), $x_i$ is a set of descriptors, and $K(x,x_i)$is the kernel function

$$K(x, x_i) = \Phi(x) \cdot \Phi(x_i) \quad (3)$$

In an SVM regression, the decision function was used in predicting or approximating the function as follows:

$$f(x) = (\sum_{i=l}^{l} \alpha_i K(x, x_i) + b) \quad (4)$$

where $\alpha_i$ is a real value, and $x_i$ is a feature vector corresponding to a training object.
Linear and non-linear regressions approximate the function by minimizing the regularized risk function $R(C)$ as follows:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \quad (5)$$

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon \text{ for } |d - y| \geq \varepsilon \\ 0 \qquad \text{otherwise} \end{cases} \quad (6)$$

where $C(1/N)\sum_{i=1}^{N} L_\varepsilon(d_i y_i)$ is the empirical error (risk) measured by the ε-insensitive loss function $L_\varepsilon(d, y)$ in which errors are not penalized below ε, and $\frac{1}{2}\|w\|^2$ is a measurement of the function flatness. The complexity ($C$) parameter is a regularized constant that is used in determining trade-off between the training error and model flatness and ε is a prescribed parameter called the tube size that approximates the accuracy placed on the training data points.

Three major learning kernels of SVM are comprised of linear, polynomial and radial basis function kernel.

Linear kernel is defined by the following equation:

$$K(x, x_i) = \Phi(x) \cdot \Phi(x_i) \quad (7)$$

where $K$ is a kernel function and Φ is a mapping function from input space onto the feature space.

Polynomial kernel is described by the following equation:

$$K(x, x_i) = \left(\langle x, x_i \rangle + 1\right)^E \quad (8)$$

where $E$ is the exponential value while a polynomial kernel with an $E$ value of 1 is essentially a linear kernel.

Radial basis function is defined by the following equation:

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \quad (9)$$

***Principal component analysis***

Principal component analysis (PCA) was performed using The Unscrambler software package, version 9.6 (Camo Software AS, Norway). Metabolic parameters were used as independent variables while the MS status was used as the dependent variable. Input variables were standardized as described by Eq. (1). The optimal number of PCs was determined according to the method of Haaland and Thomas (1988) from a plot of PCs versus the mean squared error (MSE). MSE values were calculated according to the following equation:

$$MSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n} \quad (10)$$

where $p_i$ represents the predicted output, $a_i$ represents the actual output, and $n$ represents the number of compounds presented in the data set.

***Association analysis***

Association analysis (AA) was performed using SPSS Clementine, version 11.1 (SPSS Inc., USA). AA is a data mining technique that discovers unknown relationship of items by searching for those that frequently occur together (Wang et al., 2004). The *Apriori* algorithm (Agrawal et al., 1993) is a popular method used in elucidating association rules. Support and confidence were assigned to be greater or equal

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

to the minimum values ($min_{sup}$ and $min_{conf}$) of 5 % and 80 %, respectively (Wang et al., 2004). Such parameters were used in exploring association rules. Furthermore, the *Lift* parameter was used in exploring association between various factors as well as used in the discovery of previously unknown patterns for frequently occurring health parameters of MS. Association analysis (AA) are defined as follows (Wang et al., 2004): Let $I = \{i_1, i_2, i_3, \ldots, i_m\}$ where each item represents a unique literal. A set of transaction T in a transaction database denoted by D is composed of transaction T, which contain sets of items such that $T \subseteq I$. If transaction $T$ has $X$ sets of items where $X \subseteq I$ and $Y \subseteq I$ are in a transaction $D$ where $X \cap Y = \varnothing$. Association rules are implications in the form of $X \rightarrow Y$.

The possibility of transaction $D$ in possessing X and Y is represented by the following equation:

$$\text{Support}\ (X \rightarrow Y) = \text{Support}\ (X \cup Y, D) \quad (11)$$

Furthermore, the possibility of a transaction D is composed of X also contained Y was represented in following equation:

$$\text{Confidence}\ (X \rightarrow Y) = \text{Support}\ (X \cup Y, D)/\text{Support}\ (X, D) \quad (12)$$

The *Apriori* algorithm (Agrawal et al., 1993) is a popular method used in elucidating association rules. The support and confidence were assigned to be greater than or equal to the minimum support and confidence ($min_{sup}$ and $min_{conf}$) values. In this study, association rules was obtained by using a minimum support and confidence values of $min_{sup}$= 5 % and $min_{conf}$ = 80 %, respectively. Such parameters were used to explore the association rules. Lift was employed in the association analysis as described by the following equation (Wang et al., 2004):

$$Lift = \text{Confidence}(X \rightarrow Y)/\text{Support}(Y) \quad (13)$$

***Data sampling***

Data sampling was performed by separating the data set into two subsets: (i) training set and (ii) 10-fold cross-validation (CV) testing set. 10-fold CV essentially separates the data into ten groups, leaves one group out as the testing set and uses the remaining nine groups as the training set. This process was repeated iteratively until all groups had a chance to be used as the testing set.

***Statistical analysis***

Seven statistical parameters were employed for evaluating the predictive power of the models, which is comprised of root mean squared error, sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV) (Kuo et al., 2001) and Matthews correlation coefficient (MCC) (Matthews, 1975). Equations for these statistical parameters are presented in the following equation:

Root mean square error (RMSE) was used as a measure of the predictive error of the model and was calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)}{n}} \quad (14)$$

where $p_i$, $a_i$ and $n$ are the predicted value, the actual value and the number of compounds in the data set, respectively. Ten runs of ANN calculations were performed for each investigated parameter and the average RMSE value was used for assessing the predictive performance of the model.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad (15)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (16)$$

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (17)$$

$$\text{PPV} = \frac{TP}{(TP + FP)} \quad (18)$$

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

$$\text{NPV} = \frac{TN}{(TN + FN)} \quad (19)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (20)$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives or over-predictions and *FN* is the number of false negatives or missed predictions. It should be noted that the value of MCC is 0 for a random assignment and 1.0 for a perfect prediction (Matthews, 1975).

## RESULTS

### *Population characteristics*

The study population is comprised of 15,365 participants where 6,005 (39 %) are males and 9,360 (61 %) are females. This population was previously classified into MS and non-MS groups based on a BMI cut-off of ≥ 25 kg/m$^2$ following the IDF criteria (Alberti et al., 2009). From a total of 5,638 individuals having a BMI of ≥ 25 kg/m$^2$, 2,991 had MS while 2,647 did not. The clinical and biochemical features of MS and non-MS groups in the Thai population were stratified and summarized in Table 1. These data suggests that age, BMI, SBP, DBP, FPG and TG are higher in both men and women from the MS group in comparison to those of the non-MS group. Conversely, HDL-C was lower in the MS group than in the non-MS group.

### *QPHR modeling for MS classification*

QPHR modeling is a multivariate approach for predicting MS status as a function of health parameters. Thus, the development of QPHR models essentially involves the correlation of biomedical parameters with their respective MS status. Prior to multivariate analysis, the independent variables were pre-processed to comparable scales by means of standardization using WEKA, version 3.4.5. Such standardization of variables was performed according to Eq. (1) in order to scale variables to zero mean and unit variance. In this study, several data mining techniques (i.e. DT, PCA, ANN and SVM) were employed for identifying MS in the investigated Thai population.

### *Decision tree analysis*

Decision tree or DT displayed accuracies of 99.98 % and 98.86 % for the training set and the 10-fold CV set, respectively, as shown in Table 2. The sensitivity, specificity, PPV and NPV statistical parameters demonstrated values of greater than 99 % for both the training set and 10-fold CV set (Table 2). The MCC values for the training and 10-fold CV sets were 0.9996 and 0.9972, respectively. A confusion matrix depicting the numbers of correctly and incorrectly classified individuals for the MS and non-MS groups is shown in Table 3.

### *Artificial neural network*

ANN parameters were optimized in order to obtain an optimal set of parameters. It was found that optimal values for the number of hidden node, learning epochs, learning rate and momentum are 7, 9500, 0.2 and 0.5, respectively. Statistical parameters for assessing the predictive performance of QPHR models are presented in Table 2. Accuracy and RMSE values for the training and 10-fold CV sets are 99.08 % and 0.0866, respectively, for the former and 98.78 % and 0.1005, respectively, for the latter. Sensitivity, specificity, PPV and NPV showed values greater than 98 % for both training and 10-fold CV sets (Table 2) while the MCC values were 0.9694 and 0.8391, respectively, for the training and 10-fold CV sets. Confusion matrix is presented in Table 3.

### *Support vector machine*

In order to achieve maximal performance, SVM parameters (i.e. $C$ and $\gamma$ parameters) were optimized via a two-level grid search. Initial coarse grid searches of both parameters were performed from $2^{-15}$

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

**Table 2:** A summary of statistical parameters for MS classification using decision tree analysis, artificial neural network and support vector machine

| Statistical parameters | Decision tree analysis | | Artificial neural network | | Support vector machine | |
|---|---|---|---|---|---|---|
| | Training set | 10-fold CV set | Training set | 10-fold CV set | Training set | 10-fold CV set |
| **Sensitivity** | 99.96 | 99.87 | 99.53 | 98.77 | 98.96 | 92.50 |
| **Specificity** | 99.97 | 99.85 | 98.58 | 98.79 | 97.94 | 91.40 |
| **Accuracy** | 99.98 | 99.86 | 99.08 | 98.78 | 98.47 | 91.98 |
| **PPV** | 100 | 99.87 | 98.73 | 98.93 | 98.16 | 92.38 |
| **NPV** | 99.96 | 99.85 | 99.47 | 98.60 | 98.83 | 91.54 |
| **MCC** | 0.9996 | 0.9972 | 0.9815 | 0.9754 | 0.9694 | 0.8391 |

PPV: Positive Predictive Value, NPV: Negative Predictive Value, MCC: Matthews Correlation Coefficient

**Table 3:** Confusion matrix of MS classification using decision tree, artificial neural network and support vector machine

| | Decision tree | | Artificial neural network | | Support vector machine | |
|---|---|---|---|---|---|---|
| | MS | non-MS | MS | non-MS | MS | non-MS |
| **Training set** | | | | | | |
| MS | 2990 | 1 | 2953 | 38 | 2936 | 55 |
| non-MS | 0 | 2647 | 14 | 2633 | 31 | 2616 |
| **10-fold CV** | | | | | | |
| MS | 2987 | 4 | 2959 | 32 | 2763 | 228 |
| non-MS | 4 | 2643 | 37 | 2610 | 224 | 2423 |

MS: metabolic syndrome, non-MS: non-metabolic syndrome

to $2^{15}$ using a step size of $2^2$. Results from global grid search indicated that the optimal $C$ and γ parameters were $2^{11}$ and $2^3$, respectively. Subsequently, local grid search was performed by refining the search to regions in the vicinity of the optimal values from the global grid search in the regions from $2^9$ to $2^{13}$ for the $C$ parameter while the range of $2^1$ to $2^5$ was investigated for the γ parameter using step sizes of $2^{0.25}$. Results from the local grid search indicated that the optimal values for $C$ and γ parameters were $2^{12.5}$ and $2^3$, respectively. Statistical parameters for assessing the predictive performance of the QPHR models are presented in Table 2. Good predictive performance were attained using SVM as deduced from the accuracy values for the training set and 10-fold CV set of 98.47 and 91.98, respectively, as well as from the RMS values of 0.1235 and 0.2831, respectively. Likewise, sensitivity, specificity, PPV and NPV supports this by all demonstrating values of 97 % for the training set as well as affording values greater than 91 % for the 10-fold CV set, respectively (Table 2). The MCC value also indicated good predictive performance as deduced from values of 0.9815 and 0.9754 for the training set and 10-fold CV set, respectively. Confusion matrix is displayed in Table 3.

### *Principal component analysis*

Three-dimensional displays of the PCA scores plots are shown from the 120° (Figure 1A-1D), 240° (Figure 1E-1H) and 360° point-of-views (Figure 1I-1L). These plot projects the relative distribution of data samples from the data set essentially allowing the visualization of two major clusters of data samples: the MS and non-MS clusters. For individuals having MS, Figures 1A, 1E and 1I depict the gender clusters of females and males as red and green colors, respectively, while for non-MS individuals the clusters of females and males are repre-

sented by blue and cyan colors, respectively. Figures 1B, 1F and 1J show the clustering of MS and non-MS classes in red and blue colors, respectively. In individuals having MS, Figures 1C, 1G and 1K displays the clustering of females and males in red and green colors, respectively, while for non-MS individuals, the clusters of females and males in Figures 1D, 1H and 1L displays are shown in blue and cyan colors, respectively.

### *Discovery of association rules for assessing MS*

Association analysis or AA was used in the discovery of association rules as to elucidate frequently occurring variables of metabolic abnormalities leading to MS. Binning was performed on the health parameters by transforming quantitative values to qualitative values. Particularly, binning was performed by stratifying values of the variables into several value ranges (Table 1). The binned labels for gender, age, BMI, BP, FPG, TG and HDL-C (Table 1) were used as independent variables while MS status (i.e. MS or non-MS) was used as the dependent variable. Results from AA analysis as presented in Table 4 indicated that there were a total of 43 rules for MS identification. The rule represents frequently occurring pairs of MS components in individuals having MS. It can be concluded that if an individual have any of the 43 rules (the reliability of each rule decreases as a function of frequency) then he or she are at risk of having MS. Interpretation of association rule 1 suggested that individuals having triglyceride levels of 200-299 mg/dL (TG3) along with systolic and diastolic blood pressure (BP2) of 130-139 and 85-89 mmHg, respectively, are

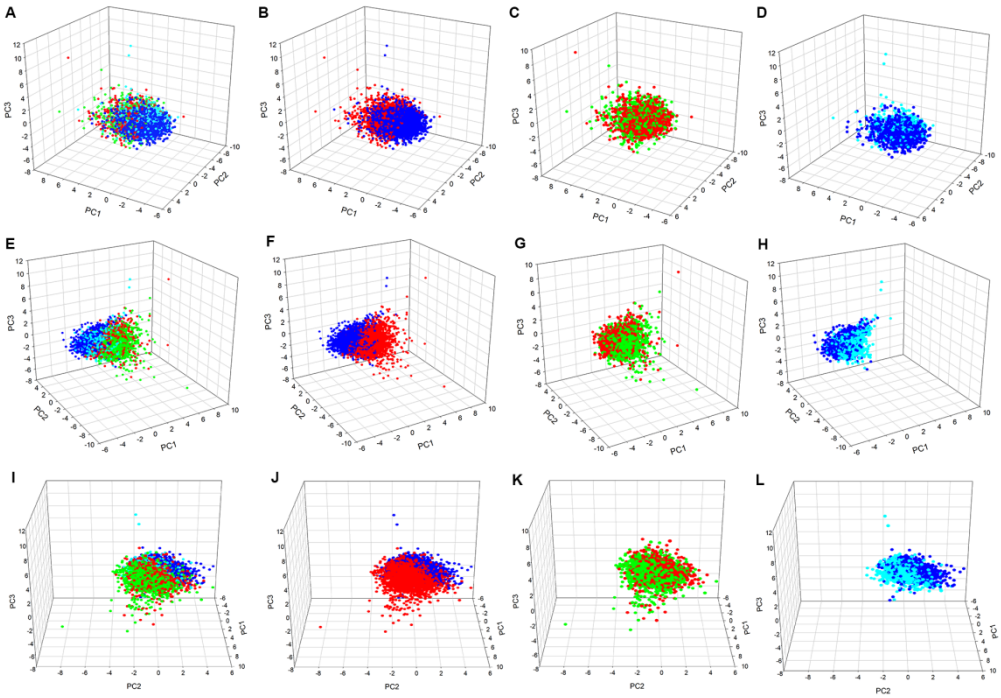


**Figure 1:** Classification of MS and non-MS using PCA as viewed from 120° (**A-D**), 240° (**E-H**) and 360°(**I-L**). **A, E** and **I** represented MS in females and males (red and green colors, respectively) and non-MS in females and males (blue and cyan colors, respectively). **B, F** and **J** represented MS as shown in red color and non-MS (blue color). **C, G** and **K** represented MS in women (red color) and in men (green color). **D, H** and **L** represented non-MS in females and males (blue and cyan colors, respectively). MS: metabolic syndrome, non-MS: non-metabolic syndrome

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

**Table 4:** Association rules for defining metabolic syndrome

| Rule ID | Association rules | | | Support (%) | Confidence (%) | *Lift* |
|---|---|---|---|---|---|---|
| | Antecedent (X) | → | Consequent (Y) | | | |
| **1** | TG3 + BP2 | → | MS | 5.29 | 100 | 1.88 |
| **2** | TG3 + FPG2 | → | MS | 5.98 | 100 | 1.88 |
| **3** | TG2 + BP3 | → | MS | 5.34 | 100 | 1.88 |
| **4** | TG2 + BP2 | → | MS | 6.42 | 100 | 1.88 |
| **5** | TG2 + FPG2 | → | MS | 7.40 | 100 | 1.88 |
| **6** | BP3 + FPG2 | → | MS | 7.64 | 100 | 1.88 |
| **7** | BP2 + FPG2 | → | MS | 9.81 | 100 | 1.88 |
| **8** | BP2 + FPG2 + TG1 | → | MS | 5.21 | 100 | 1.88 |
| **9** | HDLM2 | → | MS | 7.86 | 96.39 | 1.82 |
| **10** | HDLM2 + M | → | MS | 7.56 | 96.24 | 1.81 |
| **11** | FPG3 | → | MS | 7.68 | 93.07 | 1.75 |
| **12** | BMI6 + FPG2 | → | MS | 6.60 | 92.74 | 1.75 |
| **13** | FPG2 + Age4 | → | MS | 8.32 | 90.41 | 1.70 |
| **14** | TG3 + Age3 | → | MS | 7.96 | 88.86 | 1.68 |
| **15** | TG3 + F | → | MS | 6.10 | 87.79 | 1.65 |
| **16** | HDLF2 + Age3 + F | → | MS | 6.51 | 87.73 | 1.65 |
| **17** | HDLF2 + Age3 | → | MS | 6.62 | 87.67 | 1.65 |
| **18** | TG3 | → | MS | 17.40 | 87.05 | 1.64 |
| **19** | TG3 + M | → | MS | 10.70 | 86.73 | 1.63 |
| **20** | FPG2 + M + Age3 | → | MS | 6.94 | 86.70 | 1.63 |
| **21** | FPG2 + M | → | MS | 15.21 | 86.36 | 1.63 |
| **22** | TG2 + Age3 | → | MS | 9.93 | 85.54 | 1.61 |
| **23** | HDLF2 + F | → | MS | 14.40 | 85.47 | 1.61 |
| **24** | HDLF2 | → | MS | 14.62 | 85.32 | 1.61 |
| **25** | TG2 | → | MS | 20.95 | 84.67 | 1.60 |
| **26** | TG2 + M | → | MS | 10.41 | 84.67 | 1.60 |
| **27** | FPG2 + HDLM1 + Age3 | → | MS | 6.10 | 84.59 | 1.59 |
| **28** | TG2 + F | → | MS | 9.86 | 84.35 | 1.59 |
| **29** | FPG2 | → | MS | 30.81 | 84.34 | 1.59 |
| **30** | FPG2 + HDLM1 + M +Age3 | → | MS | 5.82 | 84.15 | 1.59 |
| **31** | FPG2 + Age3 | → | MS | 14.60 | 84.08 | 1.58 |
| **32** | FPG2 + HDLM1 | → | MS | 13.23 | 83.91 | 1.58 |
| **33** | FPG2 + HDLM1 + M | → | MS | 12.58 | 83.50 | 1.57 |
| **34** | BMI2 + FPG2 | → | MS | 5.92 | 82.04 | 1.55 |
| **35** | FPG2 + F | → | MS | 14.51 | 82.03 | 1.55 |
| **36** | TG2 + HDLM1 | → | MS | 8.92 | 81.71 | 1.54 |
| **37** | TG3 + Age2 | → | MS | 5.00 | 81.56 | 1.54 |
| **38** | TG2 + HDLM1 + M | → | MS | 8.58 | 81.40 | 1.53 |
| **39** | TG3 + HDLM1 | → | MS | 7.84 | 81.00 | 1.53 |
| **40** | FPG2 + Age3 + F | → | MS | 7.18 | 80.99 | 1.53 |
| **41** | TG2 + Age2 | → | MS | 5.21 | 80.95 | 1.53 |
| **42** | TG3 + HDLM1 + M | → | MS | 7.41 | 80.96 | 1.52 |
| **43** | BP3 + M +Age3 | → | MS | 5.04 | 80.28 | 1.51 |

Abbreviations as defined in Table 1.

EXCLI Journal 2013;12:569-583 – ISSN 1611-2156
Received: May 05, 2013, accepted: June 10, 2013, published: June 26, 2013

associated with MS. Likewise, such phenomenon can also be seen in rules 3 and 4 where individuals had abnormal triglyceride levels of 150-199 mg/dL (TG2) but displayed different ranges of blood pressure (BP2 or BP3). Furthermore, triglyceride levels of 200-299 mg/dL (TG3) and fast plasma glucose levels of 100-125 mg/dL (FPG2) in rule 2 indicated that individuals with MS also demonstrated the same patterns in rule 5 but differing in triglyceride level of 150–159 mg/dL (TG2). As for rule 11, hyperglycemic individuals having FPG greater than 126 mg/dL were correlated with MS.

## DISCUSSION

### *Predicting MS status with QPHR*

Quantitative population-health relationship or QPHR modeling is proposed herein for elucidating the relationship between biomedical parameters from individuals with respect to their metabolic syndrome status. Such QPHR model has useful implications for clinical applications in diagnosis (Firouzi et al., 2007; Kuo et al., 2001; Shin et al., 2010), health prevention (Lee et al., 2000; Nahar et al., 2011; Obenshain, 2004; Ting et al., 2009; Yoo et al., 2012) and health promotion (Lee et al., 2000; Obenshain, 2004; Ting et al., 2009; Yoo et al., 2012). A wide range of data mining techniques (i.e. ANN, SVM, MLR, PCA, DT, AA as well as self-organizing map) has previously been shown to be useful in healthcare (Lee et al., 2000; Obenshain, 2004; Ting et al., 2009; Yoo et al., 2012), medicine (Chang et al., 2011; Firouzi et al., 2007;Kim et al., 2012a, b; Nahar et al., 2011; Worachartcheewan et al., 2010a; Yeh et al., 2011), polymer chemistry (Nantasenamat et al., 2005, 2007a) and biobiological activities (Nantasenamat et al., 2007b, 2009, 2010; Thippakorn et al., 2009; Prachayasittikul et al., 2010; Worachartcheewan et al., 2011, 2012, 2013). In this study, several data mining techniques were used for assessing the MS status of an urban Thai population. The QPHR modeling approach is comparable to those of QSAR and QSPR models in which biomedical descriptors (i.e. health parameters obtained from health check-up) were subjected to multivariate analysis as to correlate them with MS status. More in-depth account of QSAR/QSPR modeling has previously been reviewed (Nantasenamat et al., 2009, 2010).

### *Comparison of the predictive performance of QPHR models*

A statistical summary of the overall predictive performance of the data mining methods employed in this study, namely DT, SVM and ANN, are presented in Table 2. The results clearly suggests that the DT model displayed the best performance as deduced from values greater than 99 % in all statistical parameters for both the training and 10-fold CV sets as well as MCC values of 0.9996 and 0.9972 for training and 10-fold CV sets, respectively. The second best performing model was those constructed by ANN, which displayed values of more than 98 % in the statistical measurements for both training and 10-fold CV sets while SVM came in last by exhibiting values of more than 97 % in the statistical parameters calculated for the training set and values greater than 91 % for the 10-fold CV set. DT has recently been used for MS identification in Thai (Worachartcheewan et al., 2010a) and Korean (Kim et al., 2012b) populations. In the study by Kim et al. (2012b) on the Korean population, it was found that TG+BP and FPG+BP were strong predictors of MS, which coincided with the results from our previous study (Worachartcheewan et al., 2010a, b) on the Thai population where such combinations of MS components were also found in MS individuals.

### *Elucidating frequently occurring pairs of MS components with AA*

AA has previously been used in clinical diagnosis for the discovery of risk factors

that are associated with the development of diseases such as diabetes (Quentin-Trautvetter et al., 2002), cancer (Nahar et al., 2011) and food borne diseases (Thakur et al., 2010). It has also been employed in identifying risk factors of occupational injury in order to prevent occupational accidences (Cheng et al., 2010; Liao et al., 2008a). Furthermore, AA is commonly applied in the business sector for the evaluation and promotion of goods in stores (Liao et al., 2008b, 2009). Findings from this study indicated that frequently occurring pairs of MS components as deduced from rules 1 through 8 were TG+BP, BP+FPG and TG+FPG. Such frequently occurring pairs were in correspondence with our previous findings (Worachartcheewan et al., 2010a, b) as well as with the investigation by Lee et al. (2008). It is interesting to note that many of the association rules are comprised of similar subset of MS components. For example, rule 7 is comprised of BP2 + FPG2 while rule 8 is comprised of BP2 + FPG2 + TG1. It can be seen that both rules contain the same subset of BP2 + FPG2 while both rules afforded the same confidence value of 100 %. In general, it was observed that association rules were essentially comprised of 2 MS components that were found to frequently occur together with factors related to gender, age and BMI. It was observed that 10 of the 43 rules were comprised of the Age3 class corresponding to the 45-54 years age group. Furthermore, 10 of the 43 rules contained the M class corresponding to males while 6 of the rules contained the F class corresponding to females. These two evidences strongly suggest that MS is associated with older age and was more prevalent in males than in females. Such results corresponded to our previous study (Worachartcheewan et al., 2010b) where the prevalence of MS in the Thai population was age-dependent and MS was found more in males than in females. Moreover, an AA study employing the *Apriori* algorithm to investigate comorbidity in Korean patients with type 2 DM, demonstrated strong association between type 2 DM and hypertension (Kim et al., 2012a). Furthermore, hypertension was found to be an important parameter associated with type 2 DM, stroke and dyslipidemia. In addition, Shin et al. (2010) employed association rule mining by means of the *Apriori* algorithm to discern association among co-morbidities of hypertension. Their results indicated that hypertension was associated with non-insulin dependent diabetes mellitus (NIDDM) and cerebral infarction. Such evidences supported our results that hypertension, dyslipidemia and hyperglycemia are important clusters of MS components that are associated with the development of CVD and type 2 DM.

## CONCLUSION

The findings strongly suggest the robustness of data mining methods (i.e. DT, ANN, SVM and PCA) for identification and classification of individuals with or without MS in an urban Thai population. The results indicated that DT was the best performing method with an accuracy of greater than 99 %. Furthermore, AA provided pertinent information on common MS components (i.e. triglyceride levels, systolic and diastolic blood pressure and fasting plasma glucose) that frequently occur together. Identification of MS components by means of association rule provided general guidelines that may potentially be used in preventing MS in individuals at risk for MS, a condition that predisposes them to the development of CVD and type 2 DM.

## ACKNOWLEDGEMENTS

A.W. is supported by the Royal Golden Jubilee (Ph.D.) scholarship of the Thailand Research Fund under the supervision of V.P and this research project is supported by the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative. We thank the Center of Medical Laboratory Services and Mobile Health Unit of the

Faculty of Medical Technology for the data set used in this study.

## REFERENCES

Agrawal R, Imilienski T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data 1993:207-16.

Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on epidemiology and prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 2009;120:1640-5.

Balkau B, Charles MA. Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). Diabet Med 1999;16:442-3.

Babu A, Fogelfeld L. Metabolic syndrome and prediabetes. Dis Mon 2006;52:55-144.

Chang C-D, Wang C-C, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011;38:5507-13.

Cheng CW, Lin C-C, Leu S-S. Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. Saf Sci 2010;48:436-44.

Cortes C, Vapnik V. Support-Vector Networks. Machine Learning 1995;20:273-97.

Firouzi F, Rashidi M, Hashemi S, Kangavari M, Bahari A, Daryani NE et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. Eur J Gastroenterol Hepatol 2007;19:1075-81.

Haaland MD, Thomas VE. Partial least squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal Chem 1988;60:1193-202.

Kim HS, Shin AM, Kim MK, Kim YN. Comorbidity study on type 2 diabetes mellitus using data mining. Korean J Intern Med 2012a;27:197-202.

Kim TN, Kim JM, Won JC, Park MS, Lee SK, Yoon SH et al. A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population. J Endocrinol Invest 2012b;35:847-52.

Kuo WJ, Chang RF, Chen DR, Lee CC. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Res Treat 2001;66: 51–7.

Lee IN, Liao SC, Embrechts M. Data mining techniques applied to medical information. Med Inform Internet Med 2000; 25:81-102.

Lee CM, Huxley RR, Woodward M, Zimmet P, Shaw J, Cho NH et al. The metabolic syndrome identifies a heterogeneous group of metabolic component combinations in the Asia-Pacific region. Diabetes Res Clin Pract 2008;81:377-80.

Liao CW, Perng YH. Data mining for occupational injuries in the Taiwan construction industry. Safety Science 2008a;46: 1091-102.

Liao S-H, Chang W-J, Lee CC. Mining marketing maps for business alliances. Expert Syst Appl 2008b;35:1338-50.

Liao S-H, Chen C-M, Hsieh C-L, Hsiao S-C. Mining information users' knowledge for one-to-one marketing on information appliance. Expert Syst Appl 2009;36:4967-79.

Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.

Nahar J, Tickle KS, Ali AB, Chen YP. Significant cancer prevention factor extraction: an association rule discovery approach. J Med Syst 2011;35:353-67.

Nantasenamat C, Naenna T, Isarankura Na Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aided Mol Des 2005;19:509-24.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Quantitative structure-imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron 2007a;22:3309-17.

Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. J Comput Chem 2007b;28:1275-89.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. EXCLI J 2009;8:74-88.

Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds Expert Opin Drug Discov 2010;5:633-54.

Nantasenamat C, Srungboonmee K, Jamsak S, Tansila N, Isarankura-Na-Ayudhya C, Prachayasittikul V. Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine. Chemometr Intell Lab Syst 2013;120:42-52.

NCEP, National Cholesterol Education Program. Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). JAMA 2001;285:2486-97.

Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004;25:690-5.

Prachayasittikul S, Wongsawatkul O, Worachartcheewan A, Nantasenamat C, Ruchirawat S, Prachayasittikul V. Elucidating the structure-activity relationships of the vasorelaxation and antioxidation properties of thionicotinic acid derivatives. Molecules 2010;15:198-214.

Quentin-Trautvetter J, Devos P, Duhamel A, Beuscart R. Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France. Stud Health Technol Inform 2002;90:557-61.

Shin AM, Lee IH, Lee GH, Park HJ, Park HS, Yoon KII et al. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc Inform Res 2010;16:77-81.

Thakur M, Olafsson S, Lee J-S, Hurburgh CR. Data mining for recognizing patterns in foodborne disease outbreaks. J Food Eng 2010;97:213-27.

Thippakorn C, Suksrichavalit T, Nantasenamat C, Tantimongcolwat T, Isarankura-Na-Ayudhya C, Naenna T et al. Modeling the LPS neutralization activity of anti-endotoxins. Molecules 2009;14:1869-88.

Ting SL, Shum CC, Kwok SK, Tsang AHC, Lee WB. Data mining in biomedicine: current applications and further directions for research. J Software Eng Appl 2009;2:150-9.

Vapnik V. Statistical learning theory. New York, USA,Wiley:1998.

Wang Y-F, Chuang Y-L, Hsu M-H, Keh H-C. A personalized recommender system for the cosmetic business. Expert Syst Appl 2004;26:427-34.

Wilson DD. Manual of laboratory & diagnostic tests. New York: McGraw-Hill, 2009.

Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques (2nd ed.). San Francisco, CA: Morgan Kaufmann, 2011.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract 2010a;90:e15-8.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Lower BMI cutoff for assessing the prevalence of metabolic syndrome in Thai population. Acta Diabetol 2010b;47(Suppl 1):91-6.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V. Predicting the free radical scavenging activity of curcumin derivatives. Chemometr Intell Lab Syst 2011;109:207-16.

Worachartcheewan A, Prachayasittikul S, Pingaew R, Nantasenamat C, Tantimongcolwat T, Ruchirawat S et al. Antioxidant, cytotoxicity and QSAR study of 1-adamantylthio derivatives of 3-picoline and phenylpyridine. Med Chem Res 2012;21: 3514–22.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V. QSAR study of amidinobis-benzimidazole derivatives as potent anti-malarial agents against Plasmodium falciparum. Chem Pap 2013, pp 1-12; DOI: 10.2478/s11696-013-0398-5.

WHO. World Health Organization consultation, definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Geneva: WHO, 1999.

WHO. Cardiovascular diseases. Geneva: WHO, 2007. http://www.who.int/mediacentre/factsheets/fs317/en/index.html (accessed 15 January 2009).

WHO. Diabetes. Geneva: WHO, 2008. http://www.who.int/mediacentre/factsheets/fs312/en/index.html (accessed 15 January 2009).

Yeh D-Y, Cheng C-H, Chen Y-W. A predictive model for cerebrovascular disease using data mining. Expert Syst Appl 2011; 38:8970-7.

Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF et al. Data Mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012; 36:2431-48.

Zupan J, Gasteiger J. Neural networks in chemistry and drug design (2nd ed.). Weinheim: Wiley-VCH, 1999.