EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

Review article:

DATA MINING FOR THE IDENTIFICATION OF METABOLIC SYNDROME STATUS

Apilak Worachartcheewan[1,2,3,*], Nalini Schaduangrat[3], Virapong Prachayasittikul[4], Chanin Nantasenamat[3]

[1] Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2] Department of Clinical Chemistry, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[3] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[4] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

[*] Corresponding author: E-mail: apilak.woa@mahidol.edu, Telephone: +66 2 441 4371 ext. 2720, Fax: +66 2 441 4380

http://dx.doi.org/10.17179/excli2017-911

ABSTRACT

Metabolic syndrome (MS) is a condition associated with metabolic abnormalities that are characterized by central obesity (e.g. waist circumference or body mass index), hypertension (e.g. systolic or diastolic blood pressure), hyperglycemia (e.g. fasting plasma glucose) and dyslipidemia (e.g. triglyceride and high-density lipoprotein cholesterol). It is also associated with the development of diabetes mellitus (DM) type 2 and cardiovascular disease (CVD). Therefore, the rapid identification of MS is required to prevent the occurrence of such diseases. Herein, we review the utilization of data mining approaches for MS identification. Furthermore, the concept of quantitative population-health relationship (QPHR) is also presented, which can be defined as the elucidation/understanding of the relationship that exists between health parameters and health status. The QPHR modeling uses data mining techniques such as artificial neural network (ANN), support vector machine (SVM), principal component analysis (PCA), decision tree (DT), random forest (RF) and association analysis (AA) for modeling and construction of predictive models for MS characterization. The DT method has been found to outperform other data mining techniques in the identification of MS status. Moreover, the AA technique has proved useful in the discovery of in-depth as well as frequently occurring health parameters that can be used for revealing the rules of MS development. This review presents the potential benefits on the applications of data mining as a rapid identification tool for classifying MS.

Keywords: metabolic syndrome, health parameters, diabetes mellitus, cardiovascular diseases, data mining, QPHR

INTRODUCTION

Over the past century, the advents in science and technology have led to significant and enormous changes in the development of countries, economies, societies and environment as well as improving quality of life. However, the effects of these advancements have led to changes and perturbation of individual/population life style, environment, culture, socioeconomic and community network. As a result, this predisposes the population with several internal and external risk factors that possibly cause pathological conditions leading up to diseases (Figure 1). These diseases occur via multiple risk factors

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

such as being infected by pathogenic microorganisms (e.g. bacteria, fungi, parasites and viruses), free radicals, carcinogens, toxic compounds, pollutants and genetic abnormalities. Moreover, lifestyle and dietary modifications as well as physical inactivity have led to metabolic abnormalities. The aforementioned risk factors possibly caused diseases such as metabolic syndrome, cardiovascular diseases, diabetes mellitus, cerebrovascular diseases, foodborne diseases, infectious diseases and cancer. Therefore, focusing on health parameters provides an interesting opportunity to explore the health status in individual and population subjects correlating with biochemical changes in the body.

Interestingly, metabolic syndrome (MS) has been implicated in the development of diabetes mellitus (DM) type 2 (WHO, 2008) and cardiovascular disease (CVD) (WHO, 2007). A MS is defined as a clustering of metabolic abnormalities, especially including central obesity (e.g. waist circumference (WC) or body mass index (BMI)), dyslipidemia (e.g. triglyceride (TG) and high-density lipoprotein-cholesterol (HDL-C)), hyperglycemia (e.g. fasting plasma glucose (FPG)), and hypertension (e.g. systolic or diastolic blood pressure (SBP or DBP)) (Alberti et al., 2009).

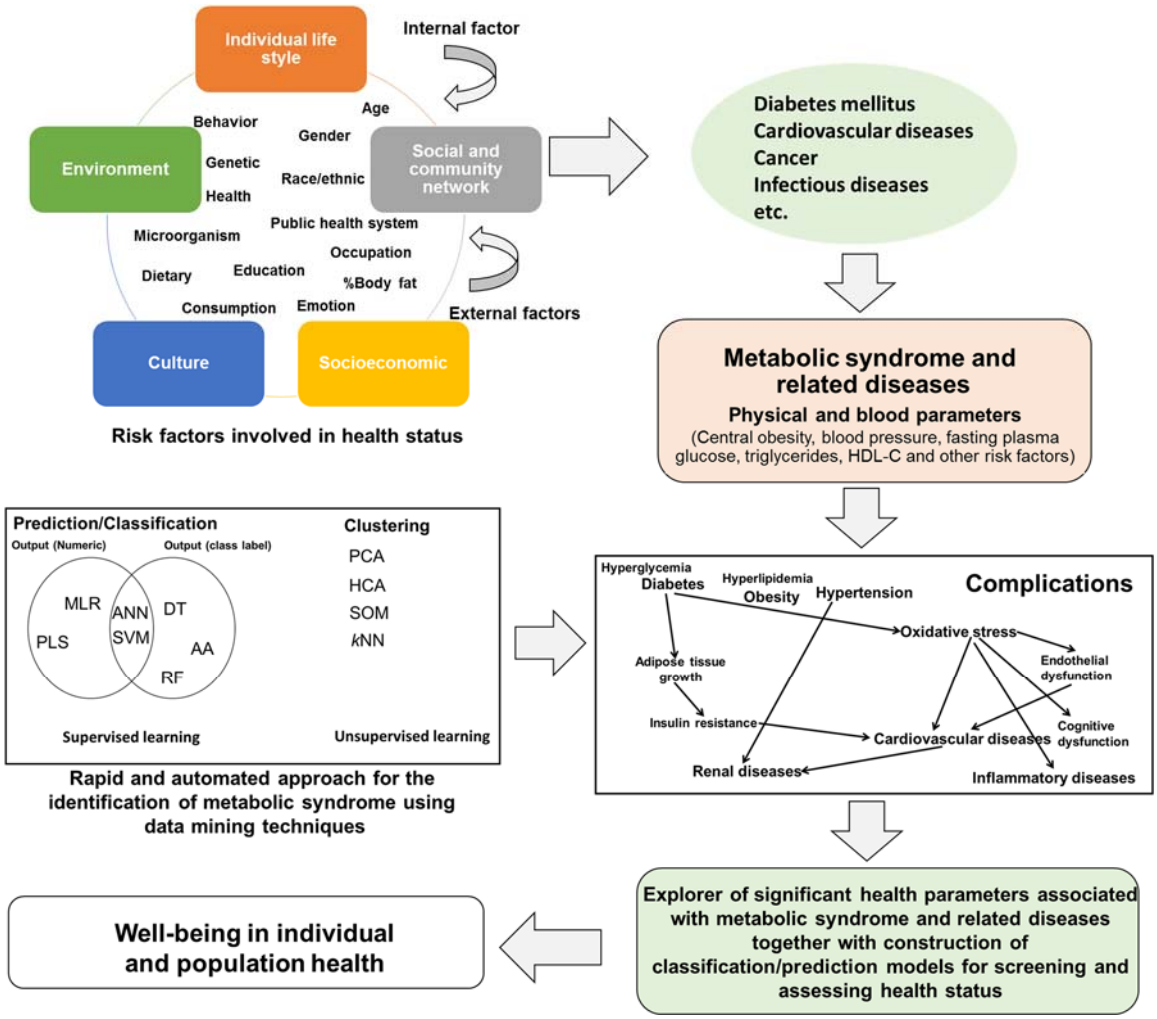

Figure 1: Risk factors of developing diseases and applications of data mining techniques for assessing health status. AA: association rule analysis, ANN: artificial neural network, DT: decision tree analysis, HCA: Hierarchical component analysis, *k*NN: *k*-nearest neighbor, MLR: multiple linear regression, PCA: principal component analysis, PLS: partial least square, RF: random forest, SOM: self-organizing map and SVM: support vector machine

The prevalence of DM has been reported in global incidences from 150 million in the year 2000 with a rapid increase to 220 million by 2010 and is estimated to reach 360 million by 2030 (Amos et al., 1997; WHO, 2008). Furthermore, the prevalence of CVD has been predicted to increase from 17.5 million in 2005 to 20 million in 2015 (WHO, 2007). Therefore, the classification of MS for rapid diagnosis to prevent the development of type 2 DM and CVD is urgently required.

The criteria for identifying MS has been developed by many organizations, for example, the first criteria was reported by the World Health Organization (WHO) in 1999 (WHO, 1999). Other criteria for defining MS have been organized by the European Group for the Study of Insulin Resistance (EGIR) (Balkau and Charles, 1999), the National Cholesterol Education Program Adult Treatment Panel III (NCEP ATPIII) (NCEP ATPIII, 2001) and the International Diabetes Federation (IDF) (Alberti et al., 2009). The criteria for identification of MS obtained from different organizations are presented in Table 1.

In fact, the geographical location, ethnicity, race as well as various social and dietary behaviors may lead to obesity, hypertension and diabetes. Moreover, according to the IDF criteria, central obesity (i.e. WC or BMI) is usually indicated as the first criteria followed by a set of two or more metabolic abnormalities. The IDF criteria uses BMI in place of waist circumference as it is significantly correlated (Ryan et al., 2008). The cut-off for obesity as outlined by the WHO is BMI $\geq$30 kg/m^2. However, this value was not appropriate for identifying the BMI status of Asian populations. This may be due to the differences in anthropometry, race/ethnic, percentage of body fat, society and dietary behaviors. Therefore, the cut-off criteria was redefined and constructed by the Steering Committee of the Regional office for the Western Pacific Region of WHO, the International Association for the Study of Obesity and the International Obesity Taskforce (WPRO) to be assigned as the new standard, whereby overweight individuals have a BMI $\geq$ 23 kg/m^2 and obese individuals have a BMI $\geq$ 25 kg/m^2 (WHO, 2000). Furthermore, Asian populations have a high record of morbidity and mortality rate arising from diabetes mellitus and cardiovascular disease even with a low threshold of central obesity with correspondingly lower waist circumference and lower BMI. Hence, the BMI cut-off for defining obesity in Asian populations was changed to 25 kg/m^2 (WHO, 2000). This new BMI cut-off has demonstrated successful identification of obesity in the Chinese (Ko et al., 2001), Japanese (Morimoto et al., 2008), Korean (Oh et al., 2004), Taiwanese (Pan et al., 2004) and Thai (Worachartcheewan et al., 2010a) populations as well as being used as the first criteria for MS identification. Furthermore, individuals with an abnormal glucose level or a corresponding insulin level as the first component of the WHO and EGIR criteria (Table 1), respectively, followed by 2 or more metabolic abnormalities were identified as having MS. In the NCEP ATPIII criteria, individuals having 3 or more of the MS components were defined as having MS (Table 1) while the IDF criteria considered the use of the central obesity as the first component followed by 2 or more abnormalities as identification for MS.

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

Table 1: Criteria for defining metabolic syndrome

CRITERIA	WHO (1999)	EGIR (1999)	NCEP ATP III (2001)	IDF (2005)
REQUIREMENT	Diabetes, impaired fasting plasma glucose, glucose intolerance or insulin resistance plus two or more of the following:	Hyperinsulinaemia (fasting insulin values above quartile for the non-diabetic population) plus with two or more of the following:	Three or more of the following:	Central obesity (ethic specific values, or BMI ≥ 30 kg/m^2) plus two or more of the following:
CENTRAL OBESITY	BMI > 30 kg/m^2 or waist-to-hip ratio > 0.90 in male or > 0.85 in female	Waist circumference ≥ 94 cm in male or ≥ 80 cm in female	Waist circumference ≥ 102 cm in male or ≥ 88 cm in female	
BLOOD PRESSURE	≥ 140/90 mmHg	≥ 140/90 mmHg) or treatment for hypertension	≥ 135/85 mmHg	≥ 135/85 mmHg or treatment for hypertension
TRIGLYCERIDE	≥ 1.7 mmol/L (150 mg/dL)	≥ 2.0 mmol/L (180 mg/dL) or treatment for dyslipidemia	≥ 1.7 mmol/L (150 mg/dL)	≥ 1.7 mmol/L (150 mg/dL) or treatment for dyslipidemia
HDL-C	< 0.9 mmol/L (35 mg/dL) in male or < 1.0 mmol/L (39 mg/dL) in female	< 1.0 mmol/L (40 mg/dL) or treatment for dyslipidemia	< 1.0 mmol/L (40 mg/dL) in male or < 1.3 mmol/L (50 mg/dL) in female	< 1.0 mmol/L (40 mg/dL) in male or < 1.3 mmol/L (50 mg/dL) in female or treatment for dyslipidemia
FASTING PLASMA GLUCOSE		≥ 6.1 mmol/L (110 mg/dL)	≥ 6.1 mmol/L (110 mg/dL)	≥ 5.6 mmol/L (100 mg/dL) or previously diagnosed Type 2 diabetes
MICROALBUMINURIA	Urinary albumin excretion rate ≥ 50 μg/min or albumin: creatinine ratio ≥ 30 mg/g	-	-	-

BMI: body mass index, EGIR: European Group for the Study of Insulin Resistance, HDL-C: High-density lipoprotein cholesterol, IDF: International Diabetes Federation, NCEP ATPIII: National Cholesterol Education Program Adult Treatment Panel III, WHO: World Health Organization

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

OVERVIEW OF DATA MINING FOR ASSESSMENT OF HEALTH STATUS

Concepts of data mining

Data mining is the process of analyzing and managing data from a large pool of information which leads to the summarization of the data for obtaining knowledge and insight into large databases which seek unknown patterns, classifications, clustering and relationships in the data set (Han and Kamber, 2001). Data mining is composed of six steps according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) established in 1996. The CRISP-DM aimed to produce a protocol on the performance of data mining that was applicable to everyone (from the novice up to an expert in the field) for a comprehensive data mining methodology and process model (Shearer, 2000). The Knowledge Discovery in Database (KDD) is also used together with data mining. The process of KDD and data mining are similar, however, data mining is one of the steps of the KDD process which includes data selection, data preprocessing, data transformation, data mining, interpretation/evaluation of the model and use of the discovered knowledge (Fayyad et al., 1996).

A typical data set as formatted in a spreadsheet or CSV text file is comprised of patients/individuals (rows) as well as health parameters and class labels (columns). Health parameters are essentially independent variables X_{1i}, X_{2i},…, X_{ni} defining the unique characteristics of patients/individuals while the class label is a dependent variable Y_i, Y_{ii},…, Y_{ni}) of each sample (Nisbet et al., 2009) as shown in Table 2.

Prior to model construction, independent variables (quantitative data) are scaled so as to afford comparison of variables by means of normalization or standardization (Nantasenamat et al., 2009, 2010).

Normalization for independent variables is adjusted in the range of 0 and 1 according to the following equation:

$$x_{ij}^{norm} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}} \qquad (1)$$

where x_{ij}^{norm} is the normalized value, x_{ij} is the value of interest, $x_j^{\min}$ is the minimum value and $x_j^{\max}$ is the maximum value.

Standardization for independent variables is performed in the mean and unit variance by using the following equation:

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

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

In addition, the original quantitative data (without normalization or standardization) and qualitative data can also be used to directly build predictive models.

In the construction of a predictive model, the data set is typically divided into two sets: 1) training set (i.e. for the training of machine learning algorithms to recognize patterns and generate models) 2) testing set (i.e. for the evaluation of the model). The types of generated testing set can be obtained from internal and external testing sets. Cross-validation is an internal testing set which divides the data set into *n* equal parts whereby one part is used as a testing set and the remaining parts are used as training sets until all parts are used as the testing set. A variety of *n*-fold cross validation selection methods have been used to evaluate the predictive models such as 10-fold cross-validation used for a large number of data set which are generated into 10 parts, for example, 500 subjects were separated into 10 equal parts, where 50 samples were used as the testing set and 450 used as the training set. In contrast, leave-one-out is employed for data sets containing a small number of objects where the numbers of folds are equal to the number of data sets. Furthermore, model validation was also performed using an external set that consists of data not used in the model construction (Nantasenamat et al., 2009, 2010).

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

Table 2: Typical data set format for data mining

Data set	Descriptor/ health parameters 1	Descriptor/ health parameters 2	Descriptor/ health parameters 3	..	Descriptor/ health parameters n	Activity/ diseases (to be modeled)
Sample i	X_{1i}	X_{2i}	X_{3i}	..	X_{ni}	Y_i
Sample ii	X_{1ii}	X_{2ii}	X_{3ii}	..	X_{nii}	Y_{ii}
Sample iii	X_{1iii}	X_{2iii}	X_{3iii}	..	X_{niii}	Y_{iii}
Sample iv	X_{1iv}	X_{2iv}	X_{3iv}	..	X_{niv}	Y_{iv}
…				..		
Sample n	X_{1n}	X_{2n}	X_{3n}	..	X_{nn}	Y_n

X represents descriptors/property/health parameters.
Y represents activity/property/diseases of interest.

The types of machine learning are categorized into 2 groups: supervised and unsupervised learning. Supervised learning consists of dependent variables assigned as numerical or class labels that make use of machine learning algorithms for the classification or prediction of the data set whereas unsupervised learning is directly performed on the data set for clustering within which dependent variables are not used (Nantasenamat et al., 2009, 2010; Nantasenamat and Prachayasittikul, 2015: Prachayasittikul et al., 2015). Examples of data mining techniques used for supervised and unsupervised learning are displayed in Figure 1. In supervised learning, the data mining techniques such as MLR, PLS, ANN and SVM are used to construct predictive models in outputs of numeric data as classification and regression models, and DT, AA, RF, ANN and SVM are used for generating classification model in output of class labels. Considering unsupervised learning, the PCA, HCA, SOM, *k*NN clustering and AA, was applied for the build-up of clustering or classifying data in unassigned output data which is used for understanding the distribution of each cluster and for identifying similar or different groups between the information. (Nantasenamat et al., 2010; Nantasenamat and Prachayasittikul, 2015; Prachayasittikul et al., 2015). Each data mining technique has shown its advantage and disadvantage such as ANN and SVM are non-linear techniques as well as black-box methods whereas MLR is an easy technique that is limited in a huge number of features. Therefore, using data mining should be considered with the type of data that can interpret significant parameters related in the output data. Furthermore, data mining could be applied in sciences and health from small molecules, chemical polymer as well as biological macromolecules up to the population level (Isarankura-Na-Ayudhya, 2009).

Data mining for medical/clinical applications

Medical/clinical databases are considered as large collections of data composed of patient/individual information such as patient history, physiological and biochemical parameters and diseases which have been collected in the hospital or laboratory systems. Therefore, understanding and revealing relationships using medical/clinical data are needed to obtain new knowledge in medical/clinical fields. Advances in the realm of computational information have allowed the development of new methods and tools for analyzing large quantities of data. Data mining has made use of medical/clinical data for discovering patterns and building predictive models (Iavindrasana et al., 2009; Koh and Tan, 2005; Lee et al., 2000; Obenshain, 2004; Ting et al., 2009; Yoo et al., 2012) to help physicians in the decision-making for diagnosis, prognosis and treatment of patients. In addition, data mining has been successfully applied for identifying and building relationship models to display the relationship between health parameters and diseases

such as cancer, cerebrovascular disease, diabetes mellitus, food-borne diseases, heart diseases, hypertension, hyperlipidemia, ischemic heart disease, inflammatory bowel disease and metabolic syndrome as shown in Table 3.

Table 3: Example of applications of data mining for medical/clinical data

Diseases/health status	Description	Data mining techniques	References
Cancer	Significant prevention factors of cancer were discovered using association rule mining algorithms.	AA	Nahar et al. (2011)
Cerebrovascular disease	Data mining was employed for constructing predictive models for cerebrovascular disease and identifying important influent factors in the disease.	DT, Bayesian classifier and Back-propagation neural network	Yeh et al. (2011)
Diabetes mellitus	AA and DT were used to analyze diabetes data from medical databases.	AA and DT	Quentin-Trautvetter et al. (2002)
	Predictive models of DM were constructed using data mining techniques (i.e. neural network, DT and logistic regression) as to select relevant factors from anthropometrical body surface scanning data.	Neural network, DT and logistic regression	Su et al. (2006)
Food-borne diseases	Patterns were identified from foodborne disease outbreaks using data mining techniques.	AA and DT	Thakur et al. (2010)
Heart diseases	Neural network and discriminant analysis methods were utilized for identifying important risk factors and high risk in heart disease patients.	Neural network and discriminant analysis	Lee et al. (2000)
Hypertension and hyperlipidemia	Different data mining techniques were applied for determining common risk factors of hypertension and hyperlipidemia.	Logistic regression, C5.0 DT, Classification and regression tree (CART), CHAID and exhaustive CHAID, Discriminant analysis and multivariate adaptive regression splines	Chang et al. (2011)
Hypertension, hyperglycemia and hyperlipidemia	Data mining techniques were used to investigate disease distribution (i.e. hypertension, hyperglycemia and hyperlipidemia) forms in various community resident areas to draw up a disease distribution map.	Hierarchical clustering	Wei et al. (2012)

Diseases/health status	Description	Data mining techniques	References
Ischemic heart disease	Back-propagation neural network and direct kernel SOM was applied to analyze magnetocardiography as to identify ischemic heart disease patients.	Back-propagation neural network and direct kernel SOM	Tantimongcolwat et al. (2008)
Inflammatory bowel disease	DT was employed for finding relevant factors and determining low bone mineral density in inflammatory bowel disease.	DT	Firouzi et al. (2007)
Metabolic syndrome	SVM and DT were used for predicting the MS	SVM and DT	Karimi-Alavijeh et al. (2016)
Metabolic syndrome	Machine learning methods (i.e. DT, ANN, SVM, PCA, AA, and RF) were utilized for classification of MS and identification of significant health parameters associated with MS.	DT, ANN, SVM, PCA, AA and RF	Worachartcheewan et al. (2010b; 2013; 2015)

AA: association rule analysis, ANN: artificial neural network, CHAID: Chi-square automatic interaction detection, DT: decision tree analysis, PCA: principal component analysis, RF: random forest, SVM: support vector machine, SOM: self-organizing map

Health parameters

Health parameters are important variables for assessing health status and for the proper diagnosis of diseases. These parameters are collected in medical databases and are obtained when individuals receive their health check-up and/or health assessment with disease conditions. Generally, a physician uses blood chemistry and physical examination together with health history and interview in order to evaluate the health status of a patient. However, delaying diagnosis of diseases may lead to morbidity and mortality for the patient. Therefore, the progression of informative computational technology can help physicians rapidly diagnose and find patterns that recognize risk factors related to developing diseases. As mention above, medical databases collecting a large amount of data are interesting and can be used as a health status evaluation of diseases for individuals. Therefore, to manage this data, powerful computational tools are necessary. Particularly, machine learning approaches namely, data mining is applied on health parameters as to discover patterns and construct predictive models of diseases. The benefit of data mining, using biomedical databases, is for the rapid and automatic diagnosis of MS in order to help with therapeutic or health prevention for individuals having risk factors for disease development.

Statistical analysis

To evaluate predictive models, the statistical parameters were performed which comprised of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) (Kuo et al., 2001) and Matthews correlation coefficient (MCC) (Matthews, 1975). These statistical parameters are calculated using the following equations:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (3)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (4)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (5)$$

$$PPV = \frac{TP}{(TP + FP)} \quad (6)$$

$$NPV = \frac{TN}{(TN + FN)} \quad (7)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

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. The value of MCC is 0 for a random assignment and 1.0 for a perfect prediction (Matthews, 1975).

QUANTITATIVE POPULATION-HEALTH RELATIONSHIP (QPHR)

The utilization of data mining techniques for assessing the health status in a population via their health parameters had previously been termed by us as quantitative population-health relationship (QPHR) (Worachartcheewan et al. 2013). QPHR makes use of data mining to elucidate the relationship between physical and biochemical parameters from populations/patients with diseases using data mining technique.

Data mining has been used to extract and explore knowledge from a large amount of data in clinical/medicinal settings. A variety of data mining techniques including SVM, ANN, MLR, PCA, SOM, DT and AA have been demonstrated for constructing predictive models of diseases (Chang et al., 2011; Firouzi et al., 2007; Kim et al., 2012; Lee et al., 2000; Nahar et al., 2011; Worachartcheewan et al., 2010a, b, 2013, 2015; Yeh et al., 2011). In addition, data mining has previously been employed to generate QSAR/QSPR models for insight into correlations between physicochemical descriptors and their biological/chemical properties (Nantasenamat et al., 2009, 2010; Nantasenamat and Prachayasittikul, 2015; Prachayasittikul et al., 2015).

QPHR models were used to discover unknown or hidden parameters associated with the progression of diseases. The QPHR models are performed with a clinical aim in diagnosis, prevention and health promotion of populations/patients. Furthermore, the QPHR models could be useful in medical/clinical data for identifying important risk factors of diseases and classifying individuals who have risk factors in development of said diseases. The procedure of QPHR is illustrated in Figure 2.

The concept of QSAR/QSPR and QPHR is similar as they are both used in the construction of predictive models for biological/chemical properties (Nantasenamat et al., 2009, 2010; Nantasenamat and Prachayasittikul, 2015) and diseases (Worachartcheewan et al., 2013), respectively. In QSAR/QSPR models, quantum chemical and molecular descriptors with their bioactivities are used to find relationships between physicochemical properties and their activities while in QPHR models, health parameters (physiological and blood chemical testing) with diseases are employed to discover patterns or elucidate the relationships between them (Table 4).

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

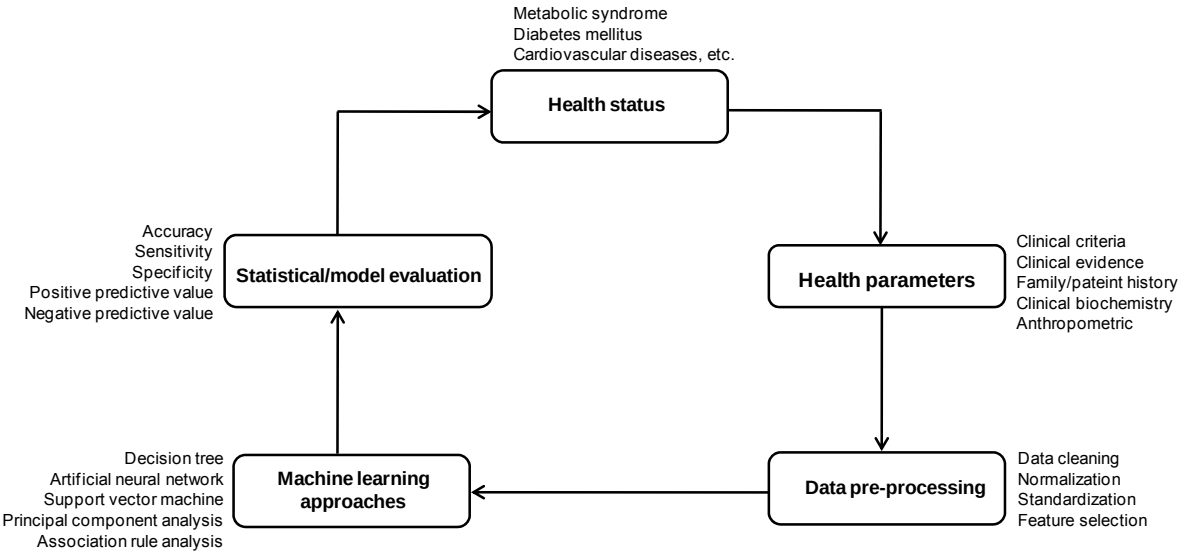

Figure 2: Schematic representation of the QPHR models

Table 4: The concept of QPHR and QSAR/QSPR models

	POPULATIONS	MOLECULES
MODEL	Quantitative Population-Health Relationship (QPHR)	Quantitative Structure-Activity/Property Relationship (QSAR/QSPR)
DESCRIPTORS	Health parameters such as physiological and blood chemistry parameters with health status	Physicochemical properties such as quantum chemical and molecular descriptors
OUTPUT	Disease or non-disease	Biological or chemical properties

The QPHR models could easily be adapted for identifying the development of other diseases. Therefore, QPHR can be used to discover unknown or hidden parameters associated with the progression of diseases for the diagnosis, prevention and health promotion in populations/patients.

In this review, examples of QPHR investigations on MS identification were described and demonstrated. In addition, Figure 1 displayed the application of data mining techniques which discovered important health parameters as well as risk factors associated with MS and related diseases together with the construction of classification/prediction models for screening and assessing health status leading to increased the well-being in individuals and population health.

MS has been focused as a risk factor associated with DM and CVD. The main cause of MS includes metabolic abnormalities in protein, carbohydrate and lipid metabolisms. Considering the MS criteria, central obesity (BMI/WC), hypertension (SBP or DBP), dyslipidemia (TG and HDL-C) and hyperglycemia (FPG) are integral component that define MS (Table 1). Furthermore, unknown components correlating with MS have been discovered whereby other factors involved in MS such as genes, socioeconomic status, behavior and dietary intake were demonstrated. In addition, the in-depth components of health parameters that occur frequently together were also illustrated using data mining. Applications of medical data mining for the classification of MS is essential for the

early detection before individuals with high risk factors develop DM and CVD.

MS classification using various machine learning approaches

Data mining has been employed to identify MS using various approaches such as ANN, SVM, RT, DT and PCA (Worachartcheewan et al., 2013; de Edelenyi et al., 2008). In addition, AA technique is also used for discovering combinations of metabolic abnormalities of MS that occur frequently together. The AA rule is correlated with the previous studies that involved metabolic abnormalities based on high levels of TG, FPG and BP and low level of HDL-C (Worachartcheewan et al., 2010a, 2010b; Lee et al., 2008). Moreover, the term of applying data mining for assessing health status via health parameters has been organized and called QPHR by Worachartcheewan et al. (2013). QPHR is defined by using health parameters for the identification associated with health status or diseases that can provide insight into the relationship between an individual's health parameters and the development of diseases. In correlating health parameters with MS status, several machine learning techniques have previously been employed, which comprises of ANN, SVM, DT and PCA (Worachartcheewan et al., 2013). Classification models for MS using various multivariate analysis have been reported that DT is the best QHPR method outperforming ANN and SVM with correct classification of MS and non-MS in greater than 99 % of cases, followed by ANN and SVM displaying an accuracy of more than 98 % and 91 % (Worachartcheewan et al., 2013), respectively. PCA is used for clustering analysis that displays distinctive MS and non-MS groups. The AA gave the rules that provide health parameters with abnormalities of MS component occurring frequently together (Worachartcheewan et al., 2013). In addition, an in-depth analysis for the identification of MS component combinations were explored using AA in order to discover metabolic abnormalities of MS components occurring frequently together. The AA was performed by stratified data from quantitative data to qualitative data using WHO and IDF criteria of metabolic abnormalities. This finding showed the combinations of MS components corresponding to previous studies and obtained association rules for the definition of MS (Worachartcheewan et al., 2010b, 2010b; Lee et al., 2008). This work was studied in the Thai population.

Interestingly, DT has been applied to find MS components in the urban and rural Korean population (Kim et al., 2012). The MS was identified using Modified National Cholesterol Education Program Adult Treatment Panel III criteria. DT displayed the combinations of high TG + high SBP, high TG + low HDL-C and high WC + high SBP + high FPG for MS in the urban population while TG + SBP + WC and SBP + WC + FPG for MS in the rural population. From this result, similar patterns for combinations of MS components were observed in the previous study and were highlighted by our results (Worachartcheewan et al., 2010a, b).

In addition, DT analysis is considered to be a robust data mining technique for constructing predictive model of metabolic syndrome status with accuracy of 73.90 % (Kim et al., 2012) and 99.86 % (Worachartcheewan et al., 2010b, 2013). Furthermore, the SVM method has been shown to yield accuracy of 75.70 % (Karimi-Alavijeh et al., 2016) and 91.98% (Worachartcheewan et al., 2013). Moreover, the CHAID decision tree has been shown to display an accuracy of 71.80 % for identifying MS. It was found that WC, TG, HDL-C, and FPG were significant health parameters for the prediction of MS (Miller et al., 2014).

The AA has been used to find patterns of MS related diseases. The study conducted on Taiwanese population by Chan et al. (2008) in MS and DM patients using AA, discovered the relationship between the diseases. It was observed that individuals having high MS were correlated with liver disease and DM individuals were associated with oral diseases such as dental carries, pulpitis, acute

gingivitis and periodontosis. Thus, the AA technique exhibited the rules of relations between diseases that can be used to help diagnosis in order to prevent illnesses in patients.

Furthermore, this technique was used to explore association rules between MS and lifestyle (Huang, 2013). It was found that individuals having a BMI >27 kg/m^2 and/or participating in vigorous physical exercise less than once a week were predisposed to having MS.

In addition, ANN and multiple logistic regression have been employed for identifying MS in patients treated with second-generation antipsychotics (SGAs) (Lin et al., 2010). The results indicated that ANN and logistic regression models gave high accuracy of 88.3 and 83.6%, respectively, while WC, BMI, DBP and gender were important variables for identifying MS in patients undergoing SGA treatment.

A study conducted on the French population by de Edelenyi et al. (2008) showed factors or combinations of factors associated with MS. Particularly, RF was applied for predicting the MS status. Dietary and genetic parameters were used as independent variables while MS or non-MS classes were used as the dependent variables. Important variables were deduced from RF including plasma concentrations of palmitoleic acid, gamma-linolenic acid (GLA) and linoleic acid. Furthermore, 3 essential single-nucleotide polymorphisms (SNPs) were selected by RF composed of APOB rs512535, LTA rs915654 and ACACB rs4766587. The correct classification is 71.4% to predict the MS status. For interpretation of health parameters, it showed that the palmitoleic acid was significantly higher in MS than non-MS while APOB rs512535 A>G and ACACB rs4766587 A>G correlated with the development of MS. Furthermore, the RF method was used to explore important health parameters and identify MS by Worachartcheewan et al. (2015). It was found that TG is considered as the first significant health parameter associated with MS and gave an accuracy > 98 % for the classification of MS. These results correlated with the previous study (Worachartcheewan et al., 2010b).

The examples of data mining application techniques are used for the classification or identification of MS. These examples help to identify patterns of MS component combinations and find the rules of metabolic abnormalities and related diseases associated with MS.

Furthermore, in this review, MS has been focused on risk factors associated with DM and CVD. MS is associated with metabolic abnormalities in protein, carbohydrate and lipid metabolisms. Concerning MS criteria, central obesity (BMI/WC), BP, dyslipidemia (TG and HDL-C) and hyperglycemia (FPG) are components that can be used to define MS. Furthermore, unknown components correlated with MS have been discovered in order to find other factors that are involved such as genes, socioeconomic status, behavior and diet. In addition, an in-depth analysis of components occurring frequently together was demonstrated via data mining. The application of medical data mining to classify MS is an essential performance for early detection of DM and CVD. Therefore, the data mining could be recommended for identification of MS during an individual's health assessment.

A summary of examples employing data mining for the classification of MS is presented in Table 5. It was used to identify patterns or combinations of MS components as well as to deduce rules for metabolic abnormalities associated with MS.

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

Table 5: Summary of identifying MS using data mining techniques

SUBJECT	HEALTH PARAMETERS	HEALTH STATUS	DATA MINING TECHNIQUE	ACCURACY (%) / ASSOCIATED FACTORS	SIGNIFICANT VARIABLES (RELATED TO MS STATUS)	REFERENCES
FRANCE	WC, FPG, HDL-C, TG, BP, dietary intake, genetic (single-nucleotide polymorphism (SNPs))	MS	RF	71.40	Palmitoleic acid of 16.1, gamma-linolenic acid (GLA), linoleic acid, APOB rs512535, LTA rs915654 and ACACB rs4766587	de Edelenyi et al. (2008)
IRAN	gender, age, weight, BMI, WC, waist-to-hip ratio (WHR), hip circumference (HC), physical activity, smoking history, hypertension, antihypertensive medication use, BP, FPG, 2-hour blood glucose, TG, total cholesterol, low-density lipoprotein cholesterol (LDL-C), HDL-C, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH)	MS	SVM DT	75.70 73.90	TG, BMI, BP, FPG TG is the most important feature for predicting MS	Karimi-Alavijeh et al. (2016)
KOREA	WC, FPG, HDL-C, TG, BP	MS	DT	Identifying urban and rural population for MS risk factors	TG, SBP, HDL-C, WC and FPG in the urban population TG, SBP, WC and FPG in the rural population TG is the most important feature for predicting MS	Kim et al. (2012)
TAIWAN	WC, FPG, HDL-C, TG, BP, diseases	MS, DM	AA	Finding common combination of disease related to MS and DM status	MS correlation with liver disease and DM associated with oral diseases	Chan et al. (2008)
TAIWAN	WC, FPG, HDL-C, TG, BP, lifestyles (i.e. shift and type of work, gender, BMI, family history, smoking, sleep quality, sleep (hrs/day), physical exercise, vigorous physical exercise (per week), alcohol, fruit intake (per week), between-meal snacks (per	MS	AA	Finding common combination of MS components	BMI, vigorous physical exercise, working in shifts, gender, coffee, family history, alcohol, physical exercise, type of work, smoking	Huang (2013)

EXCLI Journal 2018;17:72-88 – ISSN 1611-2156
Received: October 24, 2017, accepted: December 19, 2017, published: January 10, 2018

SUBJECT	HEALTH PARAMETERS	HEALTH STATUS	DATA MINING TECHNIQUE	ACCURACY (%) / ASSOCIATED FACTORS	SIGNIFICANT VARIABLES (RELATED TO MS STATUS)	REFERENCES
	week), three meals a day (per week), night snacks (per week), soft drinks with sugar (per week), coffee (per week), tea (per week)					
TAIWAN	demography (i.e. age, gender), anthropometry (i.e. BMI, BP, WC), blood chemistry (i.e. TG, FPG, HDL-C), medications (i.e. SGA agent, duration of using SGA, mood stabilizer, hypertension medication, combined antipsychotic)	MS	ANN Logistic regression	88.30 83.60	gender, WC, BMI, DBP	Lin et al. (2010)
THAILAND	demography (i.e. age, gender), anthropometry (i.e. BMI, BP), blood chemistry (i.e. FPG, blood urea nitrogen, creatinine, uric acid, cholesterol, TG, HDL-C, LDL-C, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, hemoglobin, hematocrit, white blood cell count and platelet count)	MS	ANN SVM RF DT PCA AA	98.78 91.98 98.11 99.86 Clustering Finding common combination of MS components	gender, TG, SBP, DBP, FPG and HDL-C TG is the most important feature for predicting MS	Worachartcheewan et al. (2010b, 2013, 2015)
UNITED STATES	demography (i.e. age, gender). anthropometric (i.e., weight (kg), height (cm), BMI, BP, and WC). blood chemistry (i.e. HDL-C, TG, FPG)	MS	CHAID decision trees	71.80	WC, TG, HDL-C, and FPG WC is the most important feature for predicting MS	Miller et al. (2014)

AA: association rule analysis, ANN: artificial neural network, CHAID: Chi-square automatic interaction detection, DT: decision tree analysis, MS: metabolic syndrome, PCA: principal component analysis, RF: random forest, SGAs: second-generation antipsychotics and SVM: support vector machine

CONCLUSION

This review article represents the first work of its kind whereby a summary of data mining for the assessment of MS status and discovery of in-depth MS components has been portrayed. This article summarizes the utilization of data mining techniques as a rapid identification tool for the classification of MS and non-MS categories. Complementary knowledge gained from association analysis provides pertinent information on frequently occurring parameters for defining MS. Furthermore, decision tree analysis offers insights on rules leading up to MS or non-MS groups. The topics covered in this article represent an exciting and growing area whereby various machine learning techniques offer useful insights in unravelling the mechanistic basis for MS.

The applications of data mining for the identification of MS and non-MS have been demonstrated and could potentially be employed as a rapid identification tool for classifying MS. Furthermore, association rule analysis was able to discover the important rules for defining MS. In addition, DT has been shown to be a robust machine learning approach for classifying MS and therefore holds great potential for assessing an individual's risk of MS.

Acknowledgements

This research project is supported by the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative and the research grant of Mahidol University (B.E. 2556-2558).

REFERENCES

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.

Amos AF, McCarty DJ, Zimmet P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. Diabet Med. 1997; 14(Suppl 5):S1-85.

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.

Chan C-L, Chen C-W, Liu B-J. Discovery of association rules in metabolic syndrome related diseases. In: IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence) (pp 856-62). Piscataway NJ: IEEE, 2008.

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.

de Edelenyi FS, Goumidi L, Bertrais S, Phillips C, MacManus R, Roche H, et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. Genes Nutr. 2008; 3:173–6.

Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in database. Commun ACM. 1996;39:21-6.

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.

Han J, Kamber M. Data mining: concepts and techniques. San Francisco, CA: Morgan Kaufmann Publ., 2001.

Huang YC. The application of data mining to explore association rules between metabolic syndrome and lifestyles. HIM J. 2013;42(3):29-36.

Iavindrasana J, Cohen G, Depeursinge A, Muller H, Meyer R, Geissbuhler A. Clinical data mining: a review. Yearb Med Inform. 2009:121-33.

Isarankura-Na-Ayudhya C. Protein engineering: innovation in developing biomolecules of the century. Nonthaburi, Thailand: Process Color Design & Printing Ltd Partnership, 2009.

Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler. 2016; 12:146-52.

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. 2012;35:847-52.

Ko GT, Tang J, Chan JC, Sung R, Wu MM, Wai HP, et al. Lower BMI cutoff value to define obesity in Hong Kong Chinese: an analysis based on body fat assessment by bioelectrical impedance. Br J Nutr. 2001;85:239–42.

Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag. 2005;19:64-72.

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.

Lin CC, Bai YM, Chen JY, Hwang TJ, Chen TT, Chiu HW, et al. Easy and low-cost identification of metabolic syndrome in patients treated with second-generation antipsychotics: artificial neural network and logistic regression models. J Clin Psychiatry. 2010;71:225-34.

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

Miller B, Fridline M, Liu P-Y, Marino D. Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. Comput Math Methods Med. 2014;2014:1-7.

Morimoto A, Nishimura R, Suzuki N, Matsudaira T, Taki K, Tsujino D, et al. Low prevalence of metabolic syndrome and its components in rural Japan. Tohoku J Exp Med. 2008;216:69–75.

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, Prachayasittikul V. Maximizing computational tools for successful drug discovery. Expert Opin Drug Discov. 2015; 10:321-9.

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.

NCEP ATP III. Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. 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.

Nisbet R, Elder J, Miner G. Handbook of statistical analysis & data mining application. Amsterdam: Elsevier, 2009.

Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol. 2004;5:90-5.

Oh SW, Shin SA, Yun YH, Yoo T, Huh BY. Cut-off point of BMI and obesity-related comorbidities and mortality in middle-aged Koreans. Obes Res. 2004; 12:2031–40.

Pan WH, Flegal KM, Chang HY, Yeh WT, Yeh CJ, Lee WC. Body mass index and obesity-related metabolic disorders in Taiwanese and US whites and blacks: implications for definitions of overweight and obesity for Asians. Am J Clin Nutr. 2004;79:31–9.

Prachayasittikul V, Worachartcheewan A, Songtawee N, Simeon S, Prachayasittikul V, Nantasenamat C. Computer-aided drug design of bioactive natural products. Curr Top Med Chem. 2015;15:1780-1800.

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.

Ryan MC, Fenster Farin HM, Abbasi F, Reaven GM. Comparison of waist circumference versus body mass index in diagnosing metabolic syndrome and identifying apparently healthy subjects at increased risk of cardiovascular disease. Am J Cardiol. 2008;102:40-6.

Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehouse. 2000;5:13-22.

Su C-T, Yang C-H, Hsu K-H, Chiu W-K. Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. Comput Math Appl. 2006;51:1075-92.

Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Embrechts MJ, Prachayasittikul V. Identification of ischemic heart disease via machine learning analysis on magnetocardiograms. Comput Biol Med. 2008;38 817-25.

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.

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.

Wei CK, Su S, Yang MC. Application of data mining on the development of a disease distribution map of screened community residents of Taipei county in Taiwan. J Med Syst. 2012;36:2021-7.

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

WHO. International Association for the Study of Obesity, International Obesity Taskforce. The Asia-Pacific perspective: redefining obesity and its treatment. Sydney: Health Communications, 2000.

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).

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. 2010a; 47(Suppl 1):S91-6.

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

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V**.** Quantitative population-health relationship (QPHR) for assessing metabolic syndrome. EXCLI J. 2013;12:569-83.

Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. Sci World J. 2015;2015:581501.

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.