Noname manuscript No.
(will be inserted by the editor)

Mining Health Knowledge Graph for Health Risk Prediction

**Xiaohui Tao · Thuan Pham · Ji Zhang ·
Jianming Yong · Wee Pheng Goh ·
Wenping Zhang · Yi Cai**

Received: date / Accepted: date

Abstract Nowadays classification models have been widely adopted in healthcare, aiming at supporting practitioners for disease diagnosis and human error reduction. The challenge is utilising effective methods to mine real-world data in the medical domain, as many different models have been proposed with varying results. A large number of researchers focus on the diversity problem of real-time data sets in classification models. Some previous works developed methods comprising of homogeneous graphs for knowledge representation and then knowledge discovery. However, such approaches are weak in discovering different relationships among elements. In this paper, we propose an innovative classification model for knowledge discovery from patients' personal health repositories. The model discovers medical domain knowledge from the massive data in the National Health and Nutrition Examination Survey (NHANES). The knowledge is conceptualised in a heterogeneous knowledge graph. On the basis of the model, an innovative method is developed to help uncover potential diseases suffered by people and, furthermore, to classify patients' health risk. The proposed model is evaluated by comparison to a baseline model also built on the NHANES data set in an empirical experiment. The performance of proposed model is promising. The paper makes significant contributions to the advancement of knowledge in data mining with an innovative classification model specifically crafted for domain-based data. In addition, by accessing the

X. Tao
University of Southern Queensland, Toowoomba
Tel.: +61-7-46311576
E-mail: Xiaohui.Tao@usq.edu.au

T. Pham · J. Zhang · J. Yong · W. P. Goh
University of Southern Queensland, Toowoomba

W. Zhang
Renmin University of China, China

Y. Cai
South China University of Technology, China

patterns of various observations, the research contributes to the work of practitioners by providing a multifaceted understanding of individual and public health.

Keywords Health Knowledge Graph · Classification · Healthcare · Electronic health data

1 Introduction

Improving the quality of healthcare has been one of the constant motivations to advance science and technology. The United States invested $414.3 billion in 2011 to improve the quality of healthcare systems [1]. In 2015, the United States government decided to enhance precise medicine by utilising increasingly large amounts of available health data [2]. The Australian government also launched a program recently to improve healthcare services with a ten-year plan commencing in 2015. In addition, $40 million over three years has been invested by the Australian government for indigenous primary healthcare [3], aiming at measuring and identifying effective techniques to be used in healthcare services. However, there are many challenges in developing effective models for healthcare applications. The challenges are further complicated by the volume and complexity of real data, especially in the Big Data era nowadays. Therefore, further research is in urgent demand for knowledge acquisition as well as knowledge utilisation in practical applications for healthcare services.

Knowledge discovery and knowledge management have played a very important role in healthcare services. Data mining techniques have been used to facilitate treatment evaluation and to help manage the practitioner-patient relationship. Much work has been done in disease risk assessment, with a focus on supporting medical practitioners to make safe and effective clinical decisions. Massive medical data sets contain wealthy domain knowledge that can help physicians in decision-making. Some previous research, based on the study of semantic knowledge underlying from medical corpus, has been undertaken to improve the assessment and management of disease [4, 5]. These studies have helped to improve accuracy in disease assessment and reduce errors in disease treatment. However, the diversity and complexity of terms and concepts in documents have limited the effect of study and become a significant barrier to knowledge acquisition from medical corpus.

Much effort has then been invested, trying to overcome the barrier and efficiently acquire domain knowledge underlying from medical corpus. Lee *et al.* [6] proposed a new approach to generate detailed hypotheses based on existing concepts while simultaneously using syntactic relations to support semantic representations. Some works, such as the health prediction model introduced by Chen *et al* [38], represent concepts detected from documents in a graph form. Ni *et al.* [8] proposed a new method to measure similarity among concepts. The approach represents concepts as continuous vectors, which are utilised to accumulate similar pairwise arrangements among pairs of concepts.

The study made the improvement of measuring semantics among documents successfully. Similarly, to guarantee the semantic measure of similarity between medical terms, Karpagam *et al.* [9] introduced a new method to combine different disease concepts and biomedical resources to construct a disease ontology automatically. The adoption of knowledge engineering techniques for medical domain knowledge discovery has become a promising paradigm, attracting more and more attention of the related research community.

The increasing volume of medical data is also helpful for medical professionals who wish to improve the quality of healthcare services. In relation to medical diagnosis, some researchers have developed predictive models to assess clinical risks [10,11] and to predict diseases [12,13]. However, when trying to apply these models to practice on real world data, some issues stand still. Most of data sets lack labels, whereas labelling data manually is extremely time consuming and financially expensive. The lack of features for representing all types of linked data is also an issue and putting negative effect on discovered knowledge. The real data sets are also complex in data structure and usually include multiple, heterogenous types of data. As a result, the current healthcare systems adopting data mining and machine learning techniques need to be improved for better usability of real world data.

Heterogeneous knowledge graph has then been adopted by many works while trying to mine massive, complex data for healthcare systems. Ming *et al.* [14] proposed a new algorithm to predict the label for each object by separating the different types of links and objects, which can be applied on the heterogeneous graph. The approach shows a significant improvement in the task of classification problem. Following some successful attempts, there has been an increasing number of works on classification adopting heterogeneous graph to represent data and knowledge [7,15–17]. Aiming at diagnosing disease, Chen *et al.* [38] proposed an algorithm, namely, a semi-supervised heterogeneous graph on health (SHG-Health), to predict the risk of mortality and morbidity of patients based on health examination data. The model attempts to discover knowledge from the heterogeneous graph built based on semantic relations existing in data. Their work has had a huge step up in mining health examination data for health risk prediction.

In this paper, a heterogeneous health knowledge graph is proposed to help knowledge discovery in health examination data. A real world data set, the National Health and Nutrition Examination Survey (NHANES) [1] is used in the study. The knowledge graph is constructed by a set of medical domain knowledge generated using healthcare categorization and a set of other knowledge discovered from the NHANES data set, which is a group of latent concepts decoded using *Pearson Correlation*. On the basis of the heterogeneous knowledge graph, we proposed a classification model to assess people's potential risk in having diseases. Empirical experiments have shown a promising result when compared the proposed model with Chen *et al.*'s work [38]. The contribution of our work can be highlighted as follows:

[1] https://www.cdc.gov/nchs/nhanes/index.htm

- An innovative heterogeneous knowledge graph is introduced, which is learned from a real world health examination data set.
- A classification model is proposed to predict patients' health risk using the heterogeneous knowledge graph
- A methodological contribution is delivered by the conduction of empirical experiments for evaluation of proposed models.
- The experimental results provided an evidence to support the superior of evidence-based medicine to experience-based medicine.

The study is a successful exploration of alternative mechanisms of mining medical and health data for evidence-based decision-making support.

The remainder of this paper is organised as follows. In Section 2, we review the existing work that share the same interest or on the similar research track. The research problem is then formally defined in Section 3. The proposed method is presented in Section 4, including building the health knowledge graph and adopting the knowledge graph for health risk prediction. Section 5 presents the experiment design for evaluation of the proposed classification model using health knowledge graph, and the related experimental results are reported and discussed in Section 6. Finally, Section 7 makes the conclusions.

2 Related Work

Data mining is used not only to discover latent relationships among data but also to help reveal intelligible information of users [18–20]. Data mining techniques support intelligent data preprocessing that automatically selects the required data and eliminates the undesired data, as typical works done in [21,22]. It also uses domain knowledge and automates the knowledge discovery process and leads decision-makers to a better understanding and utilization of existing knowledge in data. It is very likely that data mining could become a core technology for the practice of evidence-based medicine [23]. Since the use of computers has become extensive in the healthcare industry, data mining has also become an important modality in all fields of health sciences. The goal of research on health information is to combine computer science and information technology to improve the quality of care [24]. According to Melville *et al.* [25] and Holzinger [26], researchers use data mining as an important tool for analyzing big data to improve healthcare services. In addition, using data mining techniques, healthcare professionals can predict health insurance fraud, healthcare costs, disease prognosis, disease diagnosis and disease epidemiology and accurately estimate the length of stay (LOS) in a hospital [23].

Health status measurement prognostication has appeared as one of the most difficult challenges that health practitioners are facing. Scoring systems play an important role in minimizing errors caused by fatigue. In addition, these systems are widely used to support health practitioners in improving health knowledge and clinical decisions. In the area of myelodysplasia syndromes, different scoring systems were introduced. For example, the international prognostic scoring system (IPSS), which was introduced by Greenberg [27], focused mainly on the improvement in analyzing the specific impact of marrow blast percentage and depth of cytopenias.

The Simplified Acute Physiology Score (SAPS) II scoring system, introduced by Le Gall [28], helped physicians to make better clinical decisions by quantifying the severity of illness in the Intensive care unit area. The system introduced a method to convert the score to the possibility of a patient's mortality in the hospital. In line with SAPS II, the Acute Physiology and Chronic Health Evaluation (APACHE) II scoring system [29] was introduced, focusing on the systematic application of clinical judgments about the relative importance of derangement. The researchers have tried not only to introduce new scoring systems but also to compare the advantages and efficiency levels among the existing scoring systems, resulting in a better form of scoring systems to physicians. For example, Keegan *et al.* [30] have discussed the performance of four scoring systems including APACHE III, APACHE IV, SAPS III and Mortality Probability Model (MPM) III. The research showed that APACHE III and APACHE IV had no significant difference in distinguishing capability and they both performed better compared with SAPS III and (MPM) III. Moreover, the research also revealed that the complex models worked better than the simple models, and the efficiency level of these models depended on the number of variables.

Different studies have been conducted by using classification techniques to support health practitioners in health risk prediction. Yeh *et al.* [31] aimed to apply classification to build an optimum cerebrovascular disease predictive model. In the research, three attribute input modes, T_1, T_2, and T_3 were built, with a main focus on building efficient classification models. Alternatively, Neuvirth et al. [32] conducted research applying state-of-the-art methods to predict the health status of patients and identify potential risks. The adopted methods included logistic regression (LR) and k-nearest neighbour (KNN).

Following the similar path, a novel machine learning approach was proposed by Nguyen *et al.* [33]. The approach adopted soft labels in training process in order to refine the binary classifiers and to achieve efficient classification result. Aiming at solving the problem of label uncertainty (label noise) in binary classification, Yang *et al.* [34] introduced a new method, which is focused on using uncertain information to improve the performance of retraining-based models. The results showed that the new method is efficient and can be used to reduce human labelling errors in different applications.

The graph-based method has brought more advantages for discovering the intrinsic characteristics of data. The vertices and edges of a graph are taken up to model data points and their relationships, respectively [35]. Researchers have conducted different studies aiming to reduce the errors of the graph-based method. A study showed that data mining performance was improved significantly when the data was represented in a heterogeneous graph. Consequently, more meaningful knowledge was discovered [15, 36]. In 2010, Ming *et al.* [14] conducted a study by using a classification method, namely GNetMine, for heterogeneous networks. GNetMine used only one common classification criteria for all of the objects in the network. The common classification criteria, however, has become one of the weaknesses of the method. Alternatively, Wang *et al.* [7] argued that different types of objects in the network require different criteria of classification and introduced a new method to adopt meta paths to help mining heterogeneous graphs.

Heterogeneous information graphs have also been widely utilised in healthcare data mining, aiming at discovering meaningful medical knowledge and improving disease diagnosis. Hwang *et al.* [37] introduced a heterogeneous label propagation algorithm and adopted graph-based semi-supervised learning to discover patterns in disease genes. The study is based on homo-subnetworks where links are set up from the same type of objects to build up a heterogeneous disease-gene graph. Recently, Chen *et al.* [38] proposed an algorithm, namely semi-supervised heterogeneous graph on the health (SHG-Health), to predict high-risk disease from unlabelled data. The study made a significant contribution to healthcare data mining using heterogeneous information graph.

Data mining has made many significant contributions to discover and acquire new knowledge in various domains [19, 20]. The fact is particularly clear in healthcare domain, in which the services have been significantly improved due to the adoption of innovative data mining techniques and tools [23–26]. Specifically, many graph-based approaches have been used to assist decision making in disease diagnosis, in addition to traditional means [7, 14, 15, 35–38]). However, the majority of these approaches rely on labelled data to develop prediction algorithms, which leaves unlabeled data a fertile resource to be explored for further advancement of healthcare services. The advantages of graph-based techniques have not been universally applied in classification for health risk predictions [31–34]. In addition, the Pearson Correlation coefficient is related to the effect size of the relationship between two variables, used by Ha *et al.* to improve their model in identifying the relationship between diseases [39]. By a combination of these advantages for data mining in healthcare domain, the study presented in the paper aims to improve the understanding of semantics in healthcare data and adopts the semantic knowledge in an innovative classification model for health risk prediction.

3 Research Problem

The work is focused on patients' health risk assessment using data mining and machine learning techniques to analyse healthcare data. Our human brains have limits; medical knowledge is evolving. Healthcare practitioners may find it difficult to avoid human errors when they simply rely on their experience. Decisions made on the basis of past experience may lead to negligence of critical cases. In contrast, data mining in healthcare can help cover potentially overlooked areas because it does not have the aforementioned limitations. Data mining allows researchers to work with data collected from a huge number of patients. As a result, data mining can provide high-quality evidence covering the maximum number of possibilities to support decision-making by knowledge discovery in healthcare data.

Data analysis will help practitioners predict health risk of patients using a binary classification model. Furthermore, classification techniques not only are able to help practitioners categorise patients into broad groups such as healthy or unhealthy, but may also provide them with suggestions on what kind of diseases they are suffering from. The work presented in this paper is an attempt to answer the problem by utilising data mining and machine learning techniques. As an ultimate goal, the research aims to provide decision-making support based on evidence instead of experience, to healthcare practitioners, and to help reduce human errors.

There are several concepts and notations that need to be introduced before formally defining the research problem.

Definition 1 [Electronic Health Records]
The Electronic Health Records are a 3-tuple $\mathbb{R} := \langle \mathcal{P}, \mathcal{A}, \mathcal{M}_{\mathcal{P}}^{\mathcal{A}} \rangle$, where

- $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ is a set of patients and $|\mathcal{P}| = m$;
- $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ is a set of attributes and $|\mathcal{A}| = n$. Each attribute has a label $label(a)$ that marks the semantic meaning of a;
- $\mathcal{M}_{\mathcal{P}}^{\mathcal{A}}$ is a matrix constructed by $\mathcal{P} \times \mathcal{A}$ with values taken from a survey with questions defined by $\mathcal{A}$ for patients $\mathcal{P}$. □

Definition 2 [Patient Health Profile]
The health profile $\mathcal{HP}(p)$ of a patient $p \in \mathcal{P}$ is defined as a vector $\overrightarrow{p} = \{\langle a_1, w_1 \rangle, \langle a_2, w_2 \rangle, \ldots, \langle a_n, w_n \rangle\}$, where $a \in \mathcal{A}$ and w is the value of attribute a on patient p. □

Definition 3 [Research Problem]
Let $P = \{p_i \in \mathcal{P}, i = 1, \ldots, \imath\}$ be a set of patients; $\mathcal{S} = \{s_1, \ldots, s_K\}$ be a set of classes, where each s is a disease and K is the number of classes. Given a training set of patients $P_t = \{p_j, j = \imath + 1, \ldots, m\}$ and their respective health profiles $\mathcal{HP}(p_j)$, with $y_j^k = \{0, 1\}, k = 1, \ldots, K$ provided for describing the likelihood of p_j belonging to class s_k, the research problem is to learn a binary prediction function $f(y^k|p)$ and use it to classify $p_i \in P$ into $\{s_k\} \subset \mathcal{S}$ for prediction of the patients' health risk in terms of the set of diseases defined in $\mathcal{S}$. □

4 Health Knowledge Graph for Health Risk Classification

In this study, the National Health and Nutrition Examination Survey (NHANES) dataset is studied. NHANES is a survey conducted in the United States. The survey covers extensive topics in health and nutritional on about 10,000 people including adults and children. Thousands of questions are asked or ticked in interviews and physical examinations, resulting a total of 2585 attributes (data types) in the dataset. Eventually, the dataset contains the profile, history and health status of a large number of patients. However, in the study we focus on only adults and their health related issues, and exclude those survey data on children and those related to food and nutrition, for the sake of complexity.

Table 1 NHANES attributes by Categories

Type	Category	Attribute description
Patient Profile	***Demographics***	age, marital status, gender, education level, residential suburb, annual income, weight, people according to age groups, total number of people in the family/household, language used in interview
	Habit	consumption behavior, diet behavior and nutrition, physical activities, smoking, alcohol use, drug use
Questionnaire	***Mental Health***	questions regarding sleep disorders, depression, cognitive problems
	Current Health Status	diabetes, diagnosis of hepatitis B or hepatitis C kidney disease, sexual behaviour, osteoporosis, cardiovascular disease, dermatology, disability, immunization, oral health
	Health Conditions	asthma, childhood and adult, anaemia, psoriasis, heart, diseases, arthritis, blood transfusions
	Family History of Disease	asthma, diabetes, heart attack/angina
Examinations	***Physical***	weight, height, recumbent length, body mass, circumference muscle strength, blood pressure
	External	femur, neck, head circumference, leg, arm
	Other	trochal term, abdominal diameter, teeth, gum disease, oral hygiene, impression of soft tissue condition, denture/partial, denture/plates
Lab Tests	***Biochemical***	albumin, cholesterol, glycol haemoglobin, insulin, glucose, vitamin B12
	Blood	blood metal weights, blood lead, blood cadmium, blood mercury, blood selenium, blood manganese
	Urine	urinary arsenic, urinary creatinine, sugar, iodine, mercury, metal, urine pregnancy, trichomonas
	Other	toxocara, hepatitis, HIV antibody, human papilloma virus, nitrate, thiocyanate

Table 1 presents the information of different categories and attributes in the dataset.

4.1 Health Knowledge Graph

4.1.1 Health Knowledge Discovery from Data

The *Pearson correlation* coefficient has been commonly used in healthcare related researches to investigate the relationship between diseases [41–45]. The method identifies strong connection between factors and helps obtain optimum result in data mining and machine learning. Ha *et. al* used the *Pearson correlation* coefficient to identify the relationship between high-risk diseases and adult diseases to predict the prognosis of high risk patients [39]. Their model has achieved 78.3% accuracy compared to other classification models. Many such works have demonstrated the important effect of the *Pearson correlation* coefficient in healthcare and medical domain data mining.

The *Pearson correlation* coefficient is also adopted in our work to identify potential connection of different types of data in the dataset. The coefficient value indicates the strength of the relationship between data. Transforming each data type to a node and the connections linking these data to edges where the coefficient values indicating the strength of the links, a heterogeneous information graph can be constructed and used to help discover patterns underlying from data.

Table 2 Sample Pearson Correlation coefficient results, where the emphasised values indicate strong connection of attributes.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
a_1	**1**	**0.264**	**0.415**	**0.321**	-0.032	0.039	0.041	0	0.050
a_2	**0.264**	**1**	**0.292**	**0.401**	-0.019	0.004	0.005	0	0.021
a_3	**0.415**	**0.292**	**1**	**0.416**	-0.020	0.047	0.049	0	0.021
a_4	**0.320**	**0.401**	**0.416**	**1**	-0.049	-0.001	0	0	0.005
a_5	-0.032	-0.019	-0.020	-0.049	**1**	-0.054	-0.066	0	0.045
a_6	0.039	0.004	0.047	-0.001	-.0.054	**1**	**0.981**	0	**0.240**
a_7	0.041	0.005	0.049	0	-0.066	**0.981**	**1**	0	**0.245**
a_8	0	0	0	0	0	0	0	0	0
a_9	0.050	0.021	0.01	0.005	0.045	**0.240**	**0.245**	0	**1**

To identify the links between data and measure their strengths, the following *Pearson correlation* formula is exploited [40].

$$\upsilon = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{1}$$

where x and y are two random elements in $\mathcal{A}$. The value of the *Pearson correlation* coefficient υ reveals how one data affects the other and distinguishes the correlation between different data. With the values, those data with strong connection are clustered in a common class.

As defined in Definition 1, $\mathcal{M}_{\mathcal{P}}^{\mathcal{A}}$ is a matrix constructed by $\mathcal{A} \times \mathcal{P}$. Thus, given an attribute $a \in \mathcal{A}$, a function that returns all a's corresponding values in patients' health profiles can be defined:

$$\Omega(a_i) = \{w_i | \langle a_i, w_i \rangle \in \overrightarrow{p}, \forall p \in \mathcal{P}\} \tag{2}$$

Let $\mathcal{C}_{\mathbb{R}}$ be a set of knowledge specifying the classes revealed by correlation of the data in $\mathbb{R}$, the Electronic Health Records; c be a concept in health domain and $c \in \mathcal{C}_{\mathbb{R}}$; θ be a threshold determining if two attributes are strongly related or not. Algorithm 1 presents how the knowledge is discovered from the healthcare data set.

The health knowledge, $\mathcal{C}$, discovered by mining health data $\mathbb{R}$, is a set of health concepts as defined below:

Definition 4 [Latent Health Knowledge]
Latent health knowledge, denoted by $\mathcal{C}$, is a set of health concepts, in which

Algorithm 1: Knowledge Discovery from Health Data

```
input  : $\mathbb{R} = \{\mathcal{P}, \mathcal{A}, \mathcal{M}_{\mathcal{P}}^{\mathcal{A}}\}$;
output: $\mathcal{C}_{\mathbb{R}}$;
Let $\mathcal{C}_{\mathbb{R}} = \emptyset, isInc =$ false;
foreach $a_i \in \mathcal{A}$ do
    $\omega_i \leftarrow \Omega(a_i)$;
    foreach $a_j \in \mathcal{A}, a_i \neq a_j$ do
        $\omega_j \leftarrow \Omega(a_j)$;
        $\upsilon_{i,j} \leftarrow$ pearsonCorrelation($\omega_i, \omega_j$);
        if $\upsilon_{i,j} \geq \theta$ then
            foreach $c \in \mathcal{C}_{\mathbb{R}}$ do
                if $(a_i \in c) \wedge (a_j \notin c) \wedge (isInc =$ false$)$ then
                    $c = c \cup \{\langle (a_i, a_j), \upsilon_{i,j} \rangle\}$;
                    $isInc =$ true;
                end
                else if $(a_j \in c) \wedge (a_i \notin c) \wedge (isInc =$ false$)$ then
                    $c = c \cup \{\langle (a_i, a_j), \upsilon_{i,j} \rangle\}$;
                    $isInc =$ true;
                end
            end
            if $isInc =$ false then
                $c = \{\langle (a_i, a_j), \upsilon_{i,j} \rangle\}$;
                $\mathcal{C}_{\mathbb{R}} = \mathcal{C}_{\mathbb{R}} \cup \{c\}$;
            end
            $isInc =$ false;
        end
    end
end
return $\mathcal{C}_{\mathbb{R}}$;
```

each element is $c := \langle \mathcal{M}_A^A, \overrightarrow{\mathcal{M}_A^A} \rangle \in \mathcal{C}$, where $\mathcal{M}_A^A$ is a matric $A \times A$, $A \subset \mathcal{A}$. For each pair $(a_i, a_j) \in \mathcal{M}_A^A$, the value of $v(i,j) \in \overrightarrow{\mathcal{M}_A^A}$ indicates the strength level of correlation between a_i and a_j. □

Table 2 shows a couple of health concepts discovered by Algorithm 1. A concept comprises of attributes a_1 (MCQ160D), a_2 (MCQ160B), a_3 (MCQ160C) and a_4 (MCQ160E). They are strongly connected one another and thus, clustered in a common concept. (In fact, these attributes are commonly related to heart issues.) Another concept consists of a_6 (LBXBPB), a_7 (LBDBPBSI) and a_9 (LBXBCD). These attributes also have strong relationship and been clustered – they are actually all about blood. Attributes a_5 (WTSH2YR) and a_8 (LBDBPBLC), however, have no relationship with others listed on the table and are excluded from the "heart" and "blood" concepts. (They may belong to other classes that are not shown on the table due to the limit of space here).

4.1.2 Knowledge Acquisition in Health Domain

Domain knowledge has been widely used in data mining to help improve the performance of systems in specific domains. Xu *et al.* [46] built a model to automatically discover patterns specifying semantically similar relationships among diseases, with an aim at helping systems to access to a deeper understanding of diseases. Ni *et al.* [47] proposed an innovative method to measure similarity among concepts. Representing concepts as continuous vectors, the method accumulated pairwise similarity among pairs of concepts to measure the semantic knowledge in documents. Some other works constructed and used semantic knowledge graph to improve efficiency and (or) effectiveness in medical data analytic and data mining [48,49]. Being enlightened by their success, semantic knowledge is also adopted in our work to help discover underlying patterns from the data. A categorization in health domain is constructed based on a study of the semantic meanings of the attributes in the dataset.

Table 3 Semantic Categories

Class	**Description**
Kidney Conditions	All the attributes related to kidney disease.
Hepatitis	All types of hepatitis such as A, B, and C. In addition, some questions will be asked related to hepatitis, for example, “Have you ever received Hepatitis A vaccine?”
Diabetes	Urine or blood lab test.
Blood Pressure and Cholesterol	All the lab tests relating to blood.
Heart disease	Questions such as “Has a doctor ever told you that you had a heart attack, coronary heart disease, or congestive heart failure?” In addition, the doctor may ask about angina (angina pectoris).
Respiratory Disease	Attributes of respiratory disease, e.g., asthma, emphysema, thyroid problem, chronic bronchitis.
Profile	Personal demographics such as age, weight, and gender.
Others	Miscellanea attributes or attributes where ground truth can not be obtained.

Table 3 presents some semantic concepts with their narratives to describe the containing sub-concepts. The semantic knowledge provides a different understanding to the same data in NHANES, in addition to the health knowledge discovered by Algorithm 1 in Section 4.1.1.

The domain health knowledge, $\mathcal{S}$, acquired by categorizing the attribute labels in the health dataset $\mathbb{R}$, is a set of semantic health concepts, which is defined as:

Definition 5 [Domain Health Knowledge]
Domain Health Knowledge, denoted by $\mathcal{S}$, is a set of health concepts and their containing sub-concepts with semantic relations specified by domain experts. In the domain health knowledge

- *s is a concept containing a set of sub-concepts, $s = \{s'_1, s'_2, \ldots, s'_n\}$.*
- *s' is a sub-concept encoded from the label of an attribute, $s'_a \leftarrow label(a)$, $a \in \mathcal{A}$.*

- *$rel(x, y)$ is a Boolean function determining the existence of a relation between x and y, where $(x = s \vee s')$ and $(y = s \vee s')$.*
- *$ct(s \leftarrow s')$ indicates a "containment' relationship existing between s and s' if $rel(s, s') = 1$.*
- *$rt(s' + i \leftrightarrow s'_j)$ indicates a "related-to" relationship existing between s'_i and s'_j if $rel(s'_i, s'_j) = 1$ and $ct(s \leftarrow s'_i) \wedge ct(s \leftarrow s'_j)$.* □

From the definition, one may see that in the domain health knowledge, sub-concepts under the same concept are all related to each other, as specified by domain experts. However, such specification is not applied to the concept level. In next section, we will discuss how the relationship on the concept level is discovered using the acquired latent knowledge in Section 4.1.1.

4.1.3 Health Knowledge Graph Construction

With the knowledge discovered from the NHANES survey data and semantic knowledge acquired in health domain, we can construct a knowledge graph. The formal definition of the knowledge graph is as following:

Definition 6 [Health Knowledge Graph]
The health knowledge graph is a 2-tuple, $G :=< V, E >$ with an object mapping function $\varphi : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$, where

- *V is a set of vertices, in which each element v is a concept s or sub-concept s' in $A : \varphi(v) \in A$, $A = \mathcal{C} \cup \mathcal{S}$;*
- *E is a set of edges, in which each element is a semantic relation r in the relation type set $R : \psi(e) \in R$, where $R = \{rt, ct\}$.* □

Figure 1 illustrates a subgraph of the health knowledge graph constructed in the work. Three different health concepts are illustrated; vertice H for *Heart Disease*, P for *Patient Profile* and K for *Kidney Condition*. At the sub-concept level, vertices hs, ps and ks are shown. Concepts and sub-concepts are linked by semantic relations, where a solid line refers to *related-to* and a dash line refers to *containment*. As illustrated, *Heart Disease* (H) contains hs, *Patient Profile* (P) contains ps and *Kidney Condition* (K) contains ks. As revealed, p_1 ("age") is almost related to all sub-concept vertices in the figure because people from all ages could get a heart / kidney disease. p_2 ("height") is related to p_1 ("age") and p_3 ("weight"), however, not any sub-concepts in *Heart Disease* (H) or *Kidney Condition* (K). Also, p_3 plays an important aspect causing "heart attack" (h_1) and "weak kidney" (k_2). Simply from this simplified knowledge subgraph, one may see that *Patient Profile* (P) is related to both *Heart Disease* (H) and *Kidney Condition* (K). However, there are seems no connection between *Heart Disease* (H) and *Kidney Condition* (K) – they are independent to each other.

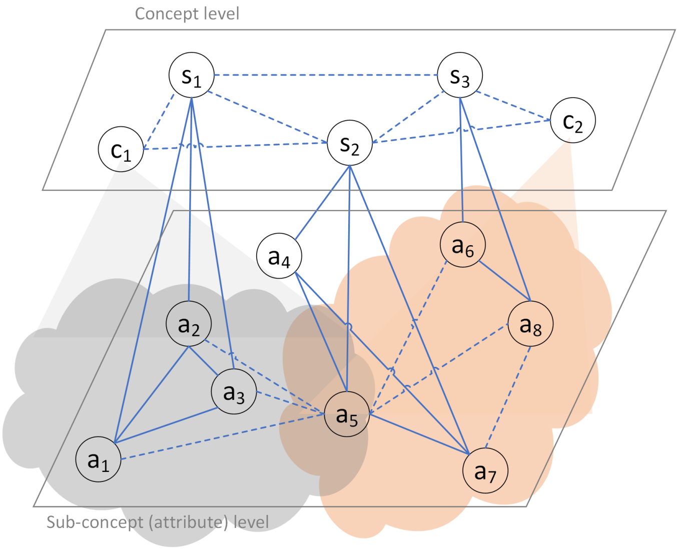

Fig. 1 A sample subgraph of the health knowledge graph

4.2 Health Knowledge Graph-based Classification

On the basis of the health knowledge graph, a function can be learned from the training data that formalises the profile of a patient for her state of healthiness or unhealthiness regarding a disease x:

$$f(x) = \sum_{i=1}^{k} v_i \times \rho(x, v_i) \times \alpha + \sum_{j=1}^{k} v_j \times \rho(x, v_i) \times \beta \tag{3}$$

where $\varphi(x) = \varphi(v_i)$ and $\varphi(x) \neq \varphi(v_j)$.

Based on $f(x)$, a patient's health status can be modelled as follows:

$$y(x) = \begin{cases} 1, \text{ if } f(x) \geq \theta \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

where θ is a threshold determining the boundary of healthiness and unhealthiness for the patient. When checking against multiple diseases $x \in \mathcal{X}$, an overall model is defined adopting Eq. 4:

$$y(\mathcal{X}) = \Pi_{n=1}^{k} y(x_n), \text{where } x \in \mathcal{X}, |\mathcal{X}| = k \tag{5}$$

In Eq. 3, α and β are two coefficients adopted to clarify the contribution of *latent health knowledge* and *domain health knowledge* in the classification model, where $\alpha + \beta = 1$. When α approximates 1 (β approximates 0), the model favourites *domain health knowledge* and omits *latent health knowledge*; when β approximates 1 (α approximates 0), *latent health knowledge* takes place and *domain health knowledge* is faded; when α and β are of the same value ($\alpha = 0.5$ and $\beta = 0.5$), *latent health knowledge* and *domain health knowledge* are equally considered in the classification model.

Aiming at finding the best values of α, β and θ, Algorithm 2 is adopted and presented as follow. The algorithm attempts to reach the best combination of α, β and θ through a convergence process for optimisation. Starting with $\theta = 0.1$, it firstly finds out the values of α and β that make the best performance of the model. Then, upon the found α and β values, the algorithm retests the model with different scales of θ and chooses the value with the greatest performance improvement. The process repeats until at a point that no more improvement could be seen, and the combination of α, β and θ is then determined.

Algorithm 2: Optimisation Algorithm

```
1: set $\theta = 0.1$;
   for each $\alpha \in \{0, 0.1, \ldots, 1\}$ where $\beta \in \{1, 0.9, \ldots, 0\}$
      – Calculate weight value for Eq.(3)
      – Select the best $\alpha$ and $\beta$
2: set the best $\alpha$ and $\beta$
   for $\theta = 0.1, 0.2, ..., 0.9$
      – Calculate the value for Eq.(3)
      – Select the best $\theta$
3: repeat step 1 for the best $\theta$
4: repeat step 2 for the best $\alpha$ and $\beta$
5: repeat step 3 and 4 until convergence
6: return $\theta$, $\alpha$ and $\beta$
```

5 Empirical Experiments for Evaluation

5.1 Experiment Design

In experiments the survey data set of NHANES was used to evaluate the proposed model and 5-*fold* method was employed to assure the reliability of evaluation result. The entire set of patient profiles in the NHANES dataset was randomly divided into five subsets, and the same experiment was conducted five rounds. In each round, one subset would be used as the testing set and the other four as the training set to construct the health knowledge graph and training the model. In the next round, another subset would be the testing set and the remaining four as the training set, and so on and son, until all five subsets were used once as the testing in a round. The final performance of the experimental models would then be the mean of the performance recorded in all five rounds.

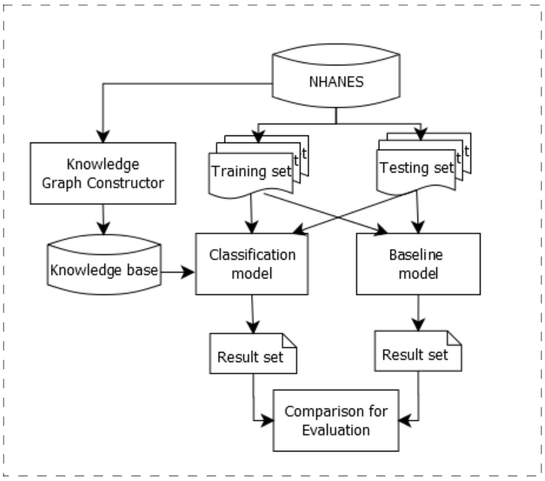

Fig. 2 Experimental Dataflow

The dataflow in experiment design is illustrated in Fig. 2. After the division of the NHANES dataset and formation of the training and testing sets, the training set data was used to mine *latent health knowledge* and *domain health knowledge*, construct the health knowledge graph, and train the classification model.The testing dataset would then be used to test the health knowledge graph-based classification model. Finally, the results were compared to the baseline model for evaluation.

5.2 Dataset

In the experiments 13 diseases were studied, as listed in Table 4. The group of patients with self-diagnosis of the diseases formed the ground truth in the testing set to evaluate the classification results produced by the experimental models. Amount the entire dataset, 4626 of 9770 participants confessed suffering from one or more diseases and were identified unhealthy. The remaining participants were considered healthy.

The data set has been prepared using data pre-processing techniques before the experiments took place. The raw dataset was highly sparse with a considerable amount of missing data. Figure 3 reports the ratio of missing data in the NHANES dataset. We chose only those attributes with substantial availability of data for the experiments. As a result, 318 out of 2585 attributes were considered in the experiments. The raw data in the dataset was also of heterogenous types, such as textual data, Boolean data, numeric data and ordinal data collected from personal demographics, observations, laboratory tests, and diagnostic reports. Data transformation and normalisation techniques were adopted to pre-process the data and made them ready for

Table 4 Studied diseases

Code	Diseases
MCQ160A	Arthritis
MCQ160L	Liver condition
HEQ030	Hepatitis C
MCQ160B	Congestive heart failure
MCQ160C	Coronary heart disease
MCQ160D	Angina, also called angina pectoris
MCQ160E	Heart attack
MCQ160G	Emphysema
MCQ160O	COPD
DIQ010	Diabetes or sugar diabetes
KIQ022	Weak or failing kidneys
BPQ080	High cholesterol
BPQ020	High blood pressure

Table 5 Statistics of the dataset

Description	Number
Participants	9770
Attributes	2858
Diseases	30
Healthy case	5144
Unhealthy case	4626

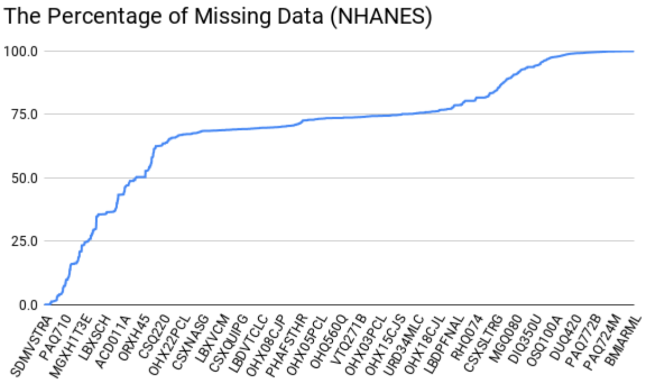

Fig. 3 The percentage of missing data

experiments. Table 5 presents the statistical information of the dataset after data pre-processing.

K-fold ($K = 5$) validation approach was adopted to help ensure the reliability of evaluation results. The NHANES dataset was randomly separated into five subsets. In each experimental run, one of the five subsets would be used for training and the other four for testing. Thirteen diseases, as shown on Table 4, were studied in experiments. The average performance of all five runs over the tests of 13 diseases then counted as the final performance of the experimental models including the proposed one and the baseline.

5.3 Baseline Model

The proposed model was evaluated by comparing with the baseline model introduced by Chen *et al* [38]. The baseline model adopted the semi-supervised learning algorithm to solve the classification problem with consideration of the relationship between different health examination data. Chen *et al.* used a general health examination (GHE) dataset collected for a group of 102,258 people living in Taipei, Taiwan from 2005 to 2010. The work categorised the 230 attributes in the dataset into three types: *physical test*, *mental assessment*, and *patient profile*. A heterogeneous graph was constructed with four different types of nodes: *Record*, *Physical Test*, *Mental Assessment*, and *Profile*. The Chen *et al*'s model was rebuilt using the NHANES dataset and compared with our proposed model in the same experimental environment, as illustrated in Fig. 2.

6 Results and Discussions

6.1 Experimental Results

The experimental models' performance was measured by the metrics of precision, recall, MAP and F_1-measure. These are modern schemes commonly used by the community for evaluation of classification models [50, 51]. They are defined as follows:

$$Percision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP is true positive (*Subject x is correctly labeled as belonging to disease y*"), TN as true negative ("*Subject x is correctly labeled as not belonging to disease y*"), FN is false negative ("*Subject x is incorrectly labeled as not belonging to disease y*"), and FP is false positive ("*Subject x is incorrectly labeled as belonging to disease y*").

Precision, also called *positive predictive value*, is the probability that subjects being labeled as belonging to a disease truly have the disease. It is calculated by dividing the number of subjects being correctly labeled as belonging to a disease by the total number of subjects labeled as belonging to the disease. Precision is the ability of a model to predict unhealthy patients. The MAP is a discriminating choice and recommended for general-purpose classification evaluation. The average precision for each disease is the mean of the precision obtained after each subject is labeled. The MAP for the 13 experimental diseases is then the mean of the average precision scores of the model in the experiments. Table 6 presents the MAP experimental results. The proposed model outperformed the baseline model with a ratio of 0.501919 vs. 0.160543.

Recall (also known as sensitivity) is the fraction that the total amount of subjects with a disease that were actually labelled with the disease by the model. It is calculated by dividing the number of subjects being correctly labeled as belonging to a disease by the total number of subjects actually belonging to the disease. Table 6 also presents the recall experimental results, where the proposed model outperformed the baseline model with a ratio of 0.602327 vs. 0.458472.

Accuracy measures the probability that subjects being labeled as belonging to a disease truly have the disease and those being excluded truly do not have the disease. It is calculated by dividing the number of subjects being correctly labeled as belonging to a disease and being correctly excluded by the total number of subjects. The accuracy result is also reported in Table 6, where the proposed model once again outperformed the baseline model by 0.855432 vs. 0.715782.

Table 6 Experimental Results, where the emphasised values indicate the superior performance in comparison.

	Proposed model	Baseline model	%Change	p-value
macro-FM	**0.547558**	0.237811	130.25%	-
micro-FM	**0.490876**	0.215951	127.31%	7.10819E-06
MAP	**0.501919**	0.160543	212.64%	9.97228E-06
Recall	**0.602327**	0.458472	31.38%	0.055774252
Accuracy	**0.855432**	0.715782	19.51%	0.053712282

Table 6 also presents the average $macro - F_1$ and $micro - F_1$ Measure results. The F_1 Measure is calculated by the following equation, where precision and recall are evenly weighted.:

$$F_1 - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{9}$$

For each topic, the $macro - F_1$ Measure averages the precision and recall and then calculates F_1 Measure, whereas the $micro - F_1$ Measure calculates the F_1 Measure for each disease and then averages the F_1 Measure values. The greater F_1 values indicate the better performance. As evidenced by the average $macro - F_1$ and $micro - F_1$ Measure results shown in Table 6, the proposed model also outperformed the based line model significantly.

Aiming at clarifying the significance of improvement achieved by the proposed model comparing with the baseline model, the percentage change in performance is used. It is calculated by the following formula, where N is the number of diseases being observed in the experiments:

$$\%Chg = \frac{1}{N} \times \sum_{i=1}^{N} \frac{result(proposed\ model) - result(baseline\ model)}{result(baseline\ model)} \times 100\% \tag{10}$$

Apparently, a larger $\%Chg$ value indicates more significant improvement achieved by the proposed model. Table 6 presents the average $\%Chg$ results achieved in experiments. As shown, the proposed model achieved significant improvements over the baseline model, especially in $macro{-}F_1$, $micro{-}F_1$ and MAP performance (marked a 130.25%, 127.31% and 212.64% improvements, respectively).

The statistical analysis using Student's Paired T-Test is also conducted, aiming at evaluating the reliability of the experimental results. The typical null hypothesis is that no difference exists in comparing two models. When two tests produce highly different significance levels (p-value <0.05), the *null* hypothesis is rejected, and the significant difference between two models can be proven. The T-Test results are also presented in Table 6. The p-values of $micro{-}F_1$ and MAP results (7.10819E-06 and 9.97228E-06, respectively) suggest that the proposed model has achieved significant improvement from the baseline model with strong rejection of *null* hypothesis. However, when looking at Recall and Accuracy results, the p-values are 0.055774252 and 0.053712282, respectively, which fail to prove the significant difference between the two models. Thus, we do not claim solid improvements made by the proposed model upon the baseline model in terms of Recall and Accuracy, though the experimental results do mark an improvement of 31.38% and 19.51%, respectively.

The detailed experimental results are reported in Table 7, where those highlighted values indicate a winning performance.

Table 7 Detailed experimental results, where the emphasised values indicate the superior performance in comparison.

Disease	Proposed model				Baseline model			
	Precision	Recall	Accuracy	F-Measure	Precision	Recall	Accuracy	F-Measure
MCQ160A	**0.5150917**	**0.7341648**	**0.7435944**	**0.6008581**	0.3393220	0.6666039	0.5692616	0.4479808
MCQ160L	**0.5895673**	0.2532995	**0.9622910**	**0.3529481**	0.0928463	**0.3437446**	0.8360992	0.1460152
HEQ030	**0.5900766**	0.6274898	**0.9925117**	**0.5920470**	0.0098732	**0.7355117**	0.3447641	0.0194795
MCQ160B	**0.2580697**	**0.6383047**	0.7591656	**0.3230750**	0.1369621	0.2317645	**0.9312353**	0.1669016
MCQ160C	**0.4652233**	**0.6137309**	**0.9560407**	**0.5282253**	0.1186306	0.2935766	0.8838613	0.1684571
MCQ160D	**0.3291395**	**0.5328138**	**0.9637438**	**0.3969341**	0.1224900	0.1474649	0.9554233	0.1282076
MCQ160E	**0.4093152**	**0.6199377**	**0.9492445**	**0.4920899**	0.1298166	0.2493632	0.9040603	0.1680419
MCQ160G	**0.3242203**	**0.8270022**	0.7841556	**0.4301281**	0.1255108	0.1116703	**0.9714428**	0.1127171
MCQ160O	**0.6661976**	**0.4533013**	0.7866677	**0.3708528**	0.1391473	0.1810321	**0.9320458**	0.1475347
DIQ010	**0.8007352**	0.5923290	**0.9570907**	**0.6779868**	0.0881521	**0.8582088**	0.3056847	0.1598454
KIQ022	**0.3991807**	**0.4302477**	0.7816202	**0.3010129**	0.0988250	0.2332426	**0.9069764**	0.1362434
BPQ020	**0.6663636**	0.7900869	**0.7955450**	**0.7223416**	0.3675907	**0.9199538**	0.4399168	0.5249895
BPQ080	**0.5117703**	0.7175411	**0.6889465**	**0.5928886**	0.3178855	**0.9880023**	0.3243942	0.4809434
Mean	**0.5019193**	**0.6023269**	**0.8554321**	**0.4908760**	0.1605425	0.4584722	0.7157820	0.2159506

Based on the experimental results, one can conclude that the proposed model is significantly better than the baseline models. These evaluation results are promising and reliable.

6.2 Discussions

The overall performance of the proposed model is better than that of the baseline model. This result suggests that the proposed model has higher capability of handling sparse data comparing with the baseline model. The training data set is sparse and non-balanced (which reflect the unreliable experimental performance comparison of two models in recall and accuracy). The proposed model achieved promising results though dealing with such sparse data. However, the baseline model was recorded with relatively lower performance, especially in MAP and F_1 measure results. The adoption of semantic and domain knowledge has made a significant impact to the success of the proposed model. The data was categorised into different categories based on the semantic and domain knowledge. The use of the *Pearson correlation* coefficient has also brought the proposed model an ability of recognising the patterns underlying from data. With all such advantages, the proposed model was leveraged and eventually outperformed the baseline significantly in overall.

Different values assigned to the coefficient (γ) in *Pearson correlation* would lead to different performance of the model. An overly-high γ value would result in meaningful information being missed in analysis. In contrast, if γ is too small, noisy data would be included and result in ambiguous analysis result. A set of empirical experiments were conducted, aiming at identifying the best value of γ in order to effectively determine if two sets of data are correlated. In the experiments, we tested the values from 0 to 1 at intervals of 0.1 for γ exhaustively. Suggested by the experimental result, $\gamma = 0.3$ gave the model the best and most stable performance, and thus was adopted.

Table 8 Sample comparison of *latent health knowledge* and *domain health knowledge* in Experiments (BPQ080)

	Threshold	Latent	Domain	Best performance			
Round	(θ)	(α)	(β)	Precision	Recall	Accuracy	F-measure
1	0.366	1	0	0.573850	0.568345	0.722309	0.571084
2	0.15	0.5	0.5	0.487437	0.738579	0.687070	0.587286
3	0.083	0.7	0.3	0.534454	0.751773	0.707055	0.624754
4	0.156	0.5	0.5	0.458333	0.759591	0.645983	0.571704
5	0.084	0.7	0.3	0.504777	0.769417	0.682316	0.609615

The coefficients of α and β in Eq. 3 were designed with an aim at leveraging the overall performance of the model by giving different considerations to the *latent health knowledge* and *domain health knowledge*. Thus, the finally determined values of α and β also reveal the importance of the *latent health* *knowledge* and *domain health knowledge* to the model. During the experiments, it was found out that, when giving the *latent health knowledge* more consideration than the *domain health knowledge* (α holds a larger value than β), the proposed model would be powered with higher performance. Table 8 presents the experimental results on *BPQ080 High cholesterol*, one of the diseases studied in the experiments, with different values setting for α, β and θ going through five rounds of K-fold. In three of five rounds, the *latent health knowledge* has presented a stronger influence than that of the *domain health knowledge* (in the second and fourth rounds they were tied). In the first round, the case is even extreme and the *domain health knowledge* appeared with no helps. Similar observation is also confirmed by experimental results on other diseases. Over all 13 studied diseases gone through five rounds each, on average α marks a value of 0.646154, which is much higher than 0.292308, the average value of β. Such an empirical result suggests that the *latent health knowledge* has played a more important role than *domain health knowledge* in the proposed model. This a true and encouraging evidence for the superior claim of "evidence-based medicine" to "experience-based medicine".

7 Conclusions and Future Work

In the recent years much effort has been invested in transforming healthcare services from traditional experience-basis to evidence basis [1–3]. Along the journey, data mining and machine learning techniques have played an important role because the *evidence* in fact refers to the latent *knowledge* discovered from massive *data* collected from daily operations of healthcare services. Data mining, machine learning and knowledge engineering / management have provided technical foundation to the transformation of healthcare services, and more advanced techniques in these areas are in great demand, aiming at further improving the quality of healthcare services.

Answering the call, we constructed a health knowledge graph in this paper using the National Health and Nutrition Examination Survey (NHANES), a health examination data set. Adopting the knowledge graph, a classification model was also introduced to predict potential health risk for patients. The *Pearson correlation* coefficient was used to discover the correlation between data attributes. Health domain knowledge contained in the categorization of diseases was also adopted in the model to help build up the knowledge graph. Aiming at evaluating the proposed classification model, empirical experiments were performed, in which the proposed model was compared with a baseline model implemented for a stat-of-the-art model introduced by Chen *et al* [38]. The experimental results showed that the proposed model outperformed the baseline model significantly in MAP and F_1 measure.

Two semantic relations, "containment" and "related-to", have been defined in Definition 5. However, distinguishing the difference in two semantic relations is not being taken as an advantage in the usage of health knowledge graph. The current design has introduced an exciting potential to our future work - to explore the influence of different semantic relations to the model and further improve the model's performance in due course. We will also endeavour to enrich the health knowledge graph using natural language processing and text mining techniques on the MEDLINE corpus, and try to expend the contributions to medical decision-support for treatment plans.

Acknowledgements The work presented in this manuscript is a significant extensive work of its preliminary form, which is published with a title "Mining Heterogeneous Information Graph for Health Status Classification" in the Proceedings of the 6th International Conference on Behavioral, Economic, and Socio-Cultural Computing, Kaohsiung, Taiwan 12–14 November 2018 (DOI: 10.1109/BESC.2018.8697292). The work is conducted with consent from the Human Research Ethics Committee of the University of Southern Queensland, Australia (Approval ID: H18REA049).

The work is partially supported by the National Natural Science Foundation of China (NSFC 71801217). The authors also like to acknowledge the use of National Health and Nutrition Examination Survey (NHAMES) in the study and specifically, thank Centers for Disease Control and Prevention of the Department of Health and Human Services, United States for making the data set publicly available for research purpose.

References

1. L. B. Mirel and K. Carper, Trends in Health Care Expenditures for the Elderly, Age 65 and Older: 2001, 2006, and 2011,2014.
2. F. S. Collins and H. Varmus, A new initiative on precision medicine, New England Journal of Medicine, vol. 372, no. 9, pp. 793–795, 2015.
3. K. Gardner, B. Sibthorpe, M. Chan, G. Sargent, M. Dowden, and D. McAullay, Implementation of continuous quality improvement in Aboriginal and Torres Strait Islander primary health care in Australia: a scoping systematic review, BMC health services research, vol. 18, no. 1, p. 541, 2018.
4. Y.-T. Cheng, Y.-F. Lin, K.-H. Chiang, and V. S. Tseng, Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease, IEEE journal of biomedical and health informatics, vol. 21, no. 2, pp. 303–311, 2017..
5. C. Y. Chin, M. Y. Weng, T. C. Lin, S. Y. Cheng, Y. H. K. Yang, and V. S. Tseng, Mining disease risk patterns from nationwide clinical databases for the assessment of early rheumatoid arthritis risk, PloS one, vol. 10, no. 4, e0122508, 2015.
6. Lee, JB., Kim, J and Park, JC 2006, Automatic extension of Gene Ontology with flexible identification of candidate terms? Bioinformatics, Oxford University Press, Vol. 22 No. 6, pp. 665–670.
7. C. Luo, R. Guan, Z. Wang, and C. Lin, Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks, in European Conference on Information Retrieval, 2014, pp. 210–221.
8. Y. Ni, QK. Xu, F. Cao, Y. Mass, D. Sheinwald, H. J. Zhu, and S. S. Cao 2016, Semantic Documents Relatedness using Concept Graph Representation, in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16, ACM Press, New York, New York, USA, pp. 635–644
9. P. Karpagam, S. Sivasubramanian, and C. Nalini, Extending Disease Ontology with Newly Evaluated Terms to Improve Semantic Medical Information Retrieval, International Journal of Applied Engineering Research, vol. 11, no. 5, pp. 3527–3535, 2016.
10. J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, Adaptive mining prediction model for content recommendation to coronary heart disease patients, Cluster computing, vol. 17, no. 3, pp. 881–891, 2014.
11. M. Sabibullah, V. Shanmugasundaram, and R. Priya, Diabetes patients risk through soft computing model, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 2, no. 6, pp. 60–65, 2013.

12. C.-D. Chang, C.-C. Wang, and B. C. Jiang, Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors, Expert systems with applications, vol. 38, no. 5, pp. 5507–5513, 2011.
13. F. Huang, S. Wang, and C.-C. Chan, Predicting disease by using data mining based on healthcare information system, in 2012 IEEE International Conference on Granular Computing, 2012, pp. 191–194.
14. M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, Graph regularized transductive classification on heterogeneous information networks, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010, pp. 570–586.
15. Y. Sun, Y. Yu, and J. Han, Ranking-based clustering of heterogeneous information networks with star network schema, in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 797–806.
16. M. Wan, Y. Ouyang, L. Kaplan, and J. Han, Graph regularized meta-path based transductive regression in heterogeneous information network,in Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 918–926.
17. X. Kong, P. S. Yu, Y. Ding, and D. J. Wild, Meta path-based collective classification in heterogeneous information networks, in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, pp. 1567–1571.
18. H. Xie, Q. Li, X. Mao, X. Li, Y. Cai and C. Zheng, Mining Latent User Community for Tag-Based and Content-based Search in Social Media. The Computer Journal, 57(9), 1415–1430, 2014.
19. D. J. Hand, Principles of data mining, Drug safety, vol. 30, no. 7, pp. 621–622, 2007.
20. H. C. Koh, G. Tan, and others, Data mining applications in healthcare, Journal of healthcare information management, vol. 19, no. 2, p. 65, 2011.
21. H. Xie, Q. Li, X. Mao, X. Li, Y. Cai and Y. Rao, Community-aware User Profile Enrichment in Folksonomy. Neural Networks, v58, pp. 111–121, 2014.
22. H. Xie, X. Li, T. Wang, L. Chen, K. Li, F. L. Wang, Y. Cai, Q. Li and H. Min, Personalized Search for Social Media via Dominating Verbal Context. Neurocomputing, 172(C), 27–37, 2016.
23. I. Yoo et al., data mining in healthcare and biomedicine: a survey of the literature, Journal of medical systems, vol. 36, no. 4, pp. 2431–2448, 2012.
24. M. Herland, T. M. Khoshgoftaar, and R. Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.
25. S. Rosset, C. Perlich, G. Swirszcz, P. Melville, and Y. Liu, Medical data mining: insights from winning two competitions, Data Mining and Knowledge Discovery, vol. 20, no. 3, pp. 439–468, 2010.
26. A. Holzinger, Machine learning for health informatics, in Machine Learning for Health Informatics, Springer, 2016, pp. 1–4.
27. P. L. Greenberg et al., Revised international prognostic scoring system (IPSS-R) for myelodysplastic syndromes, Blood, p. blood012, 2012.
28. P. Prakash, K. Krishna, and D. Bhatia, Usefulness of SAPS II scoring system as an early predictor of outcome in ICU patients, J Indian Acad Clin Med, vol. 7, no. 3, pp. 202–5, 2006.
29. D. P. Wagner and E. A. Draper, Acute physiology and chronic health evaluation (APACHE II) and Medicare reimbursement, Health care financing review, vol. 1984, no. Suppl, p. 91, 1984.
30. M. T. Keegan, O. Gajic, and B. Afessa, Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance, Chest, vol. 142, no. 4, pp. 851–858, 2012.
31. D.-Y. Yeh, C.-H. Cheng, and Y.-W. Chen, A predictive model for cerebrovascular disease using data mining, Expert Systems with Applications, vol. 38, no. 7, pp. 8970–8977, 2011.
32. H. Neuvirth et al., Toward personalized care management of patients at risk: the diabetes case study, in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 395–403.
33. Q. Nguyen, H. Valizadegan, and M. Hauskrecht, Learning classification models with soft-label information, Journal of the American Medical Informatics Association, vol. 21, no. 3, pp. 501–508, 2014.

34. Y. Yang and M. Loog, Active learning using uncertainty information, in Pattern Recognition (ICPR), 2016 23rd International Conference on, 2016, pp. 2646–2651.
35. A. Guillory and J. A. Bilmes, Label selection on graphs, in Advances in Neural Information Processing Systems, 2009, pp. 691–699.
36. B. Long, Z. M. Zhang, X. Wu, and P. S. Yu, Spectral clustering for multi-type relational data, in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 585–592.
37. T. Hwang and R. Kuang, A heterogeneous label propagation algorithm for disease gene discovery, in Proceedings of the 2010 SIAM International Conference on Data Mining, 2010, pp. 583–594.
38. L. Chen, X. Li, Q. Z. Sheng,W.-C. Peng, J. Bennett, H-Y. Hu and N. Huang, Mining Health Examination Record: Graph-Based Approach, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9, pp. 2423–2437, 2016.
39. J.-W. Ha et al., Predicting high-risk prognosis from diagnostic histories of adult disease patients via deep recurrent neural networks, in Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on, 2017, pp. 394–399.
40. L. Egghe and L. Leydesdorff,The relation between Pearson's correlation coefficient r and Salton's cosine measure, Journal of the American Society for information Science and Technology, vol. 60, no. 5, pp. 1027–1036, 2009.
41. X. Zhou, J. Menche, A.-L. Barabsi, and A. Sharma, Human symptoms disease network, Nature communications, vol. 5, p. 4212, 2014.
42. A. Tsanas, M. A. Little, and P. E. Mcsharry, A methodology for the analysis of medical data, in Handbook of Systems and Complexity in Health, Springer, pp. 113–125, 2013.
43. S. O. Torres, H. Eicher-Miller, C. Boushey, D. Ebert, and R. Maciejewski, Applied Visual Analytics for Exploring the National Health and Nutrition Examination Survey, 2012 45th Hawaii Int. Conf. Syst. Sci., pp. 1855–1863, 2012.
44. H. Al-Mubaid and H. A. Nguyen, Measuring semantic similarity between biomedical concepts within multiple ontologies, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 39, no. 4, pp. 389–398, 2009.
45. I. Alonso and D. Contreras,Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach, Expert Systems with Applications, vol. 44, pp. 386–399, 2016.
46. R. Xu, L. Li, and Q. Wang, RiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text, BMC bioinformatics, vol. 15, no. 1, p. 105, 2014.
47. Y. Ni, QK. Xu, F. Cao, Y. Mass, D. Sheinwald, HJ. Zhu, and SS. Cao, Semantic documents relatedness using concept graph representation, in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016, pp. 635–644.
48. L. Diem, J.-P. Chevallet, and D. T. B. Thuy, Thesaurus-based query and document expansion in conceptual indexing with UMLS, in Research, Innovation and Vision for the Future, 2007 IEEE International Conference on (2008), 2007.
49. A. B. Abacha and P. Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, Journal of biomedical semantics, vol. 2, no. 5, p. S4, 2011.
50. D. Bowes, T. Hall, and D. Gray, Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix, in Proceedings of the 8th International Conference on Predictive Models in Software Engineering, 2012, pp. 109–118.
51. G. P. Visa and P. Salembier, Precision-recall-classification evaluation framework: Application to depth estimation on single images, in European Conference on Computer Vision, 2014, pp. 648–662.