Mining Health Examination Records — A Graph-based Approach

Ling Chen, Xue Li, *Member, IEEE,* Quan Z. Sheng, *Member, IEEE,* Wen-Chih Peng, *Member, IEEE,* John Bennett, Hsiao-Yun Hu, and Nicole Huang

Abstract—General health examination is an integral part of healthcare in many countries. Identifying the participants at risk is important for early warning and preventive intervention. The fundamental challenge of learning a classification model for risk prediction lies in the unlabeled data that constitutes the majority of the collected dataset. Particularly, the unlabeled data describes the participants in health examinations whose health conditions can vary greatly from healthy to very-ill. There is no ground truth for differentiating their states of health. In this paper, we propose a graph-based, semi-supervised learning algorithm called SHG-Health (Semi-supervised Heterogeneous Graph on Health) for risk predictions to classify a progressively developing situation with the majority of the data unlabeled. An efficient iterative algorithm is designed and the proof of convergence is given. Extensive experiments based on both real health examination datasets and synthetic datasets are performed to show the effectiveness and efficiency of our method.

Index Terms—Health examination records, semi-supervised learning, heterogeneous graph extraction.

1 Introduction

Huge amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction [1], [2], [3], [4], [5], [6], [7], [8], [9]. An EHR contains digitally stored healthcare information about an individual, such as observations, laboratory tests, diagnostic reports, medications, procedures, patient identifying information, and allergies [10]. A special type of EHR is the Health Examination Records (HER) from annual general health check-ups. For example, governments such as Australia, U.K., and Taiwan [11], [12], [13], offer periodic geriatric health examinations as an integral part of their aged care programs. Since clinical care often has a specific problem in mind, at a point in time, only a limited and often small set of measures considered necessary are collected and stored in a person's EHR. By contrast, HERs are collected for regular surveillance and preventive purposes, covering a comprehensive set of general health measures (for example, see Table 1), all collected at a point in time in a systematic way [14]. Identifying participants at risk based on their current and past HERs is important for early warning and preventive intervention. By "risk", we mean unwanted outcomes such as mortality and morbidity.

In this study we formulated the task of risk prediction as a multi-class classification problem using the Cause of Death (COD) information as labels, regarding the health-related death as the "highest risk". The goal of risk prediction is to effectively classify 1) whether a health examination participant is at risk, and if yes, 2) predict what the key associated disease category is. In other words, a good risk prediction model should be able to exclude low-risk situations and clearly identify the high-risk situations that are related to some specific diseases.

A fundamental challenge is the large quantity of unlabeled data. For example, 92.6% of the 102,258 participants in our geriatric health examination dataset do not have a COD label. The semantics of such "alive" cases can vary from generally healthy to seriously ill, or anywhere in between. In other words, there is no ground truth available for the "healthy" cases. If we simply treat this set of alive cases as the negative class, it would be a highly noisy majority class. On the other hand, if we take this large alive set as *genuinely unlabeled*, as opposed to cases with known labels removed, it would become a multi-class learning problem with large unlabeled data.

Most existing classification methods on healthcare data do not consider the issue of unlabeled data. They either have expert-defined low-risk or control classes [1], [2], [3], [4], [5], [6] or simply treat non-positive cases as negative [7], [8], [9]. Methods that consider unlabeled data [15], [16], [17], [18], [19], [20], [21], [22], [23] are generally based on Semi-Supervised Learning (SSL) [24] that learns from both labeled and unlabeled data. Amongst these SSL methods, only [18], [19] handle large and genuinely unlabeled health data. However, unlike our scenario, both methods are designed for binary classification and have predefined negative cases. A closely related approach is Positive and Unlabeled (PU) learning [25], [26], [27], which can be seen as a special case of SSL with only positive labels available. While the unlabeled

- *L. Chen and X. Li are with School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia. E-mail: l.chen5@uq.edu.au,xueli@itee.uq.edu.au*
- *Q.Z. Sheng is with School of Computer Science, The University of Adelaide, Adelaide, Australia. E-mail: michael.sheng@adelaide.edu.au*
- *W-C. Peng is with Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan. E-mail: wcpeng@cs.nctu.edu.tw*
- *J. Bennett is with University Health Service, The University of Queensland, Brisbane, Australia. E-mail: john.bennett@uq.edu.au*
- *H-Y. Hu is with Department of Education and Research, Taipei City Hospital, Taipei, Taiwan. E-mail: A3547@tpech.gov.tw*
- *N. Huang is with Institute of Hospital and Health Care Administration, National Yang-Ming University, Taipei, Taiwan. E-mail: syhuang@ym.edu.tw*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2561278, IEEE Transactions on Knowledge and Data Engineering

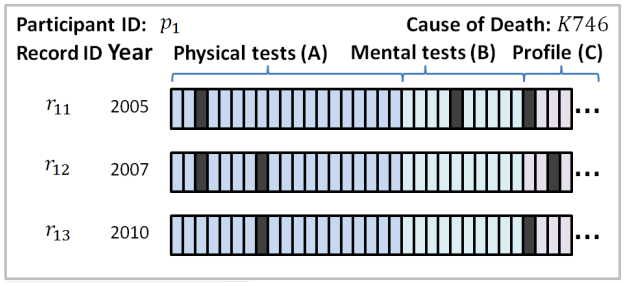

Fig. 1: An example of health examination records of participant p_1 who took examinations in three non-consecutive years, 2005, 2007, and 2010. Test items are in different categories (A,B,...) and the abnormal results are marked black. The main cause of death of p_1 was cirrhosis of liver encoded as $K746$.

set U in a PU learning problem is similar to our alive set, its existing applications in healthcare only address binary classification problem. Nguyen *et al.* introduced a multi-class extension called mPUL [28]; however, their method used a combined set of negative and unlabeled example, while in our case negative example is not available.

The other key challenge of HERs is *heterogeneity*. Fig. 1 demonstrates the health examination records of Participant p_1 in three non-consecutive years with test items in different categories (e.g., physical tests, mental tests, etc.) and abnormal results marked black. This example shows that 1) a participant may have a sequence of irregularly time-stamped longitudinal records, each of which is likely to be sparse in terms of abnormal results, and 2) test items are naturally in categories, each conveying different semantics and possibly contributing differently in risk identification. Therefore this heterogeneity should be respected in the modeling.

This paper proposes a semi-supervised heterogeneous graph-based algorithm called **SHG-Health** (Semi-supervised Heterogeneous Graph on Health) as an evidence-based risk prediction approach to mining longitudinal health examination records. To handle *heterogeneity*, it explores a *Hetero*geneous graph based on Health Examination Records called **HeteroHER** graph, where examination items in different categories are modelled as different types of nodes and their temporal relationships as links. To tackle large *unlabeled data*, SHG-Health features a semi-supervised learning method that utilizes both labeled and unlabeled instances. In addition, it is able to learn an additional $K+1$ "unknown" class for the participants who do not belong to the K known high-risk disease classes.

The main contributions of this work are three-fold:

- We present the SHG-Health algorithm to handle a challenging multi-class classification problem with substantial unlabeled cases which may or may not belong to the known classes. This work pioneers in risk prediction based on health examination records in the presence of large unlabeled data.
- A novel graph extraction mechanism is introduced for handling heterogeneity found in longitudinal health examination records.
- The proposed graph-based semi-supervised learning algorithm SHG-Health that combines the advantages from heterogeneous graph learning and class discovery shows significant performance gain on a large and comprehensive real health examination dataset of 102,258 participants as well as synthetic datasets.

The rest of the paper is organized as follows. Section 2 reviews existing works on mining health examination data and learning methods that handle unlabeled health data, followed by Section 3 where the background on graph-based semi-supervised learning is discussed. Section 4 presents the proposed SHG-Health algorithm for evidence-based risk prediction. In Section 5, we demonstrate the effectiveness and efficiency of our proposed algorithm based on both real datasets and synthetic datasets. Section 6 concludes our work and discusses future research directions.

2 Related Work

In this section we review existing related studies, namely those on mining health examination data and those on classification with unlabeled data in healthcare applications.

2.1 Data Mining on Health Examination Records

Although Electronic Health Records (EHRs) have attracted increasing research attention in the data mining and machine learning communities in recent years [2], [3], [5], [6], [8], [9], [29], mining general health examination data is an area that has not yet been well-explored, except a few studies on risk prediction such as the chronic disease early warning system proposed in [30] and our previous work on health score classification framework [8], [31]. However, none of the them considered unlabeled data. In addition, the approach presented in [8] is limited to a binary classification problem (using alive/deceased labels) and consequently it is not informative about the specific disease area in which a person is at risk. The existing studies on healthcare data that handled unlabeled data are discussed in the next subsection.

2.2 Classification with Unlabeled Healthcare Data

Unlabeled data classification are commonly handled via *Semi-Supervised Learning* (SSL) that learns from both labeled and unlabeled data [24], and *Positive and Unlabeled (PU) learning*, a special case of SSL that learns from positive and unlabeled data alone [25].

PU learning [25] is often adapted for disease gene classification when only the labels for disease genes are available [26], [27]. Recently, Nguyen *et al.* proposed *mPUL* [28], a multi-class PU learning model for activity recognition. The method trains m 1-vs-others binary probabilistic base classifiers, each trained with a positive set and a merged set of negative and unlabeled instances. The class decision is based on the maximum class probability greater than 0.5; otherwise the unknown class is predicted. However, it is not directly applicable to our problem, since we do not have negative instances available for training.

SSL has attracted increasing attention in healthcare applications based on EHRs [16], [17], [18], [19], [20], [21], [22], [23]. At the molecular level, Huang *et al.* proposed iSELF [17], a SSL method based on local Fisher discrimination analysis for disease gene classification. Nguyen *et al.* [18]

constructed a protein-protein interaction network, which defines interacted genes as candidate genes and the rest as negative genes for SSL based on Gaussian fields and harmonic functions [24]. At the disease level, many graph-based SSL were proposed. Garla *et al.* [19] applied Laplacian SVM as a SSL approach for cancer case management. Wang *et al.* [20] proposed a graph-based SSL method that is able to learn patient risk groups for patient risk stratification. Kim *et al.* [16] proposed a co-training graph-based SSL method for breast cancer survivability prediction. It iteratively assigns pseudo-labels to unlabeled data when there is a consensus amongst the learners and includes the pseudo-labeled instances in the labeled set until the unlabeled set stops decreasing. Zhang *et al.* [23] constructed a bipartile graph for ranking-based lung nodule image classification. Liu *et al.* [22] constructed a temporal graph based on event sequence for temporal phenotyping. However, different from our case, none of these methods consider an "unknown" class and they all have predefined instances for all classes, either by experts [16], [17], [19], [20], [22], [23] or via other mechanisms [18]. In addition, unlike our approach, all the graph-based SSL methods above used homogeneous graphs.

3 Background

Learning from labelled and unlabeled data is often called semi-supervised learning or transductive inference [32]. Graph-based methods that model data points as vertices and their relationships as edges on graph, are often used to exploit the intrinsic characteristics of data [33]. Zhu *et al.* [24] proposed an algorithm based on Gaussian fields and harmonic functions to propagate labels to the unlabeled data, which can be interpreted as a random walk on graph. Zhou *et al.* [32] introduced the Learning with Local and Global Consistency (LLGC) algorithm that spreads the label information of each point to its neighbors to achieve both local and global consistency. A graph can be constructed either 1) based on real-world networked data [34], [35], [36], such as from social networks, bibliographic networks, and webpage networks, or 2) by computing affinity matrices to encode the similarity between data points [32], [37]. Many *graph-based semi-supervised learning* (GSSL) methods can be viewed as estimating a function of soft labels F based on two assumptions on graph [20], [24], [32], [37], [38]. The *smoothness* assumption states that F should not change much for nearby points, and the *fitness* assumption requires that F should not change much from the ground-truth labels. By adapting a graph-based approach and exploring the underlying graph structure of health examination records with semi-supervised learning, our method is capable of handling large unlabeled data.

To further tackle the issues of the absence of ground truth for the "healthy" cases and the heterogeneity embedded in the examination records, we utilized class discovery methods to handle the "unknown" class and heterogeneous graph representations for GSSL as follows.

3.1 Class Discovery for GSSL

Situations arise when unlabeled data may belong to unknown or latent classes. Nie *et al.* [37] introduced a scholastic graph-based semi-supervised learning (GGSSL) method for novel class discovery (if the number of classes is known) or outlier detection (if otherwise). By introducing an instance-level parameter α that assigns little weight to unlabeled data and large weight to the labeled data, GGSSL allows the soft label scores of unlabeled vertices on the graph to be updated according to their connectivities to labeled vertices. Wang *et al.* [20] further modified the model to discover more than one unseen class for patient risk stratification based on a patient graph constructed using ICD codes. Recently Zhao *et al.* [38] extended GGSSL for classification on Alzheimer's Disease, by introducing a compact graph construction strategy via minimizing local reconstruction error. However, all of the above algorithms are limited to homogeneous graphs, where vertices belong to one object type, and thus are by themselves not capable of handling the heterogeneity embedded in health examination records.

To train a disease risk prediction model that is capable of identifying high-risk individuals given no ground truth for "healthy" cases, we treated the "unknown" class as a class to be learned from data. We incorporated the class discovery mechanism of [37] into our method to handle the "unknown" class.

3.2 Heterogeneous GSSL

Traditional GSSL methods are limited to homogeneous graphs [16], [19], [20], [32], [37], [38]. However, it has been recognized in recent years that networks of heterogeneous types of objects are prevalent in the real world [21], [34], [35], [39]. For example in healthcare applications, methods that explore the heterogeneous structure of gene-phenotype networks have been developed [21], [40]. The term "network medicine" [39] has been coined to refer to a broad approach to human disease based on a complex intracellular and intercellular network that connects tissue and organ systems.

For the heterogeneous extensions of GSSL algorithms, Hwang *et al.* [21] proposed a heterogeneous label propagation algorithm based on GSSL for disease gene discovery. Their heterogeneous disease-gene graph was constructed based on homo-subnetworks that link same-type objects together and the mutual interactions between homo-subnetworks. The algorithm iteratively propagates the label scores via homo-subnetworks and hetero-subnetworks until convergence. Ji *et al.* proposed GNetMine [35] to work on a heterogeneous graph of multi-type objects, known as a heterogeneous information network [34]. The classification process can be intuitively viewed as a process of knowledge propagation throughout the network across different types of objects through links. GNetMine was originally designed for bibliographic information networks that are intrinsically heterogeneous and was shown to outperform other GSSL methods with homogeneous graphs. They further proposed RankClass [36] based on the same framework with additional updates on the local weighted graph for individual classes. However, the above methods were designed for a multi-class semi-supervised learning problem with predefined classes, and thus have no mechanism for handling the "unknown" class. Inspired by GNetMine and RankClass, we integrated a heterogeneous component into our method to handle heterogeneity.

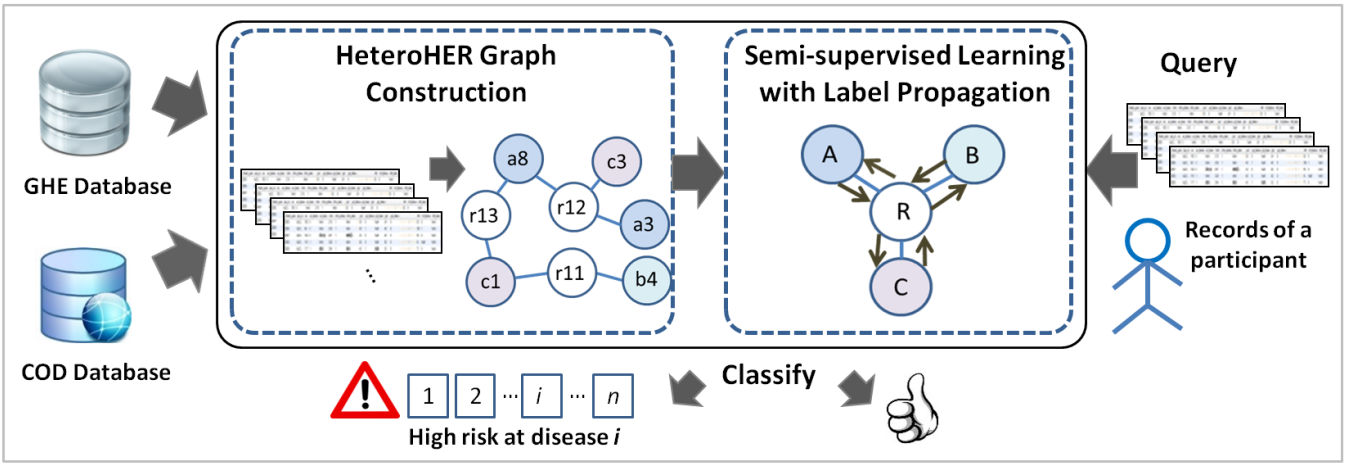

Fig. 2: An overview of the proposed SHG-Health algorithm for risk prediction.

In summary, our proposed SHG-Health algorithm can be seen as combining the advantages of GGSSL [37] and GNetMine [35] for solving a practical clinical problem of risk prediction from longitudinal health examination data with heterogeneity and large unlabeled data issues.

4 SHG-Health

To solve the problem of health risk prediction based on health examination records with heterogeneity and large unlabeled data issues, we present a semi-supervised heterogeneous graph-based algorithm called SHG-Health. The semi-supervised learning problem is formulated as follows:

Problem Definition 1. Given a set of health examination records of n participants $S = \{s_1, ..., s_l, s_{l+1}, ..., s_n\}$, where $s_i = \{r_{i1}, ..., r_{in_i}\}$ is the set of n_i records of participant i and r_{ij} is a tuple (x_{ij}, t_{ij}) such that $x_{ij} \in \mathbb{R}^d$ is a d-dimensional vector for the observations at time t_{ij}, and a set of labels $C = \{1, ..., c\}$, the first l participants s_i $(i \leq l)$ are labeled as $y_i \in C$ and the remaining $u = n - l$ participants $s_{l+1}, ..., s_{l+u}$ are unlabeled ($l \ll u$). The goal is to predict for unlabeled $s_i (l < i \leq n)$ a label $y_i \in \tilde{C} = \{1, ..., c, c+1\}$ where $c+1$ gives a mechanism to handle an additional class for unknown cases.

An overview of our proposed solution to the problem is included in Fig. 2, above. Our SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) labels described in Section 5.1 as inputs. Its key components are a process of *Hetero*geneous Health Examination Record (HeteroHER) graph construction and a semi-supervised learning mechanism with label propagation for model training. Given the records of a participant p_i as a query, SHG-Health predicts whether p_i falls into any of the high-risk disease categories or "unknown" class whose instances do not share the key traits of the known instances belonging to a high-risk disease class.

4.1 HeteroHER Graph

A graph representation allows us to model data that is sparse. To capture the heterogeneity naturally found in health examination items, we constructed a graph called HeteroHER consisting of multi-type nodes based on health examination records.

4.1.1 Graph Construction

The process of HeteroHER graph construction includes the following steps:

Step 1. **Binarization:** As a preparatory step, all the record values are first discretized and converted into a $0/1$ binary representation, which serves as a vector of indicators for the absence/presence of a discretized value. Specifically, real values, such as age, are first binned into fixed intervals (e.g., 5 years). Then, all the ordinal and categorical values are converted into binary representations.

Step 2. **Node Insertion:** Every element in the binary representation obtained in Step 1 with a value "1" is modeled as a node in our HeteroHER graph, except that only the abnormal results are modeled for examination items (both physical and mental). This setting is primarily based on the observation that physicians make clinical judgements generally based on the reported symptoms and observed signs, and secondarily for the reduction of graph density.

Step 3. **Node Typing:** Every node is typed according to the examination category that its original value belongs to, for example, the Physical tests (A), Mental tests (B), and Profile (C) in Fig. 1. In addition, a new type of nodes is introduced to represent individual records such as r_{11}, r_{12}, and r_{13} in the same figure. All the other non-Record type nodes that are linked to the Record type nodes can be seen as the *attribute* nodes of these Record type nodes. In other words, categories A, B, and C in Fig. 1 can be regarded as the attributes of the Record type at a schema level. This leads to a graph schema with a star shape as shown on the right of Fig. 3 below, which is known as a star schema [34]. Note that types can often be hierarchically structured and thus choosing the granularity of node type may require domain knowledge or be done experimentally.

Step 4. **Link Insertion:** Every attribute (non-Record) type node is linked to a Record type node representing the record that the observation was originally from. The weight of the links is calculated based on the assumption that the newer a record the more important it is in terms

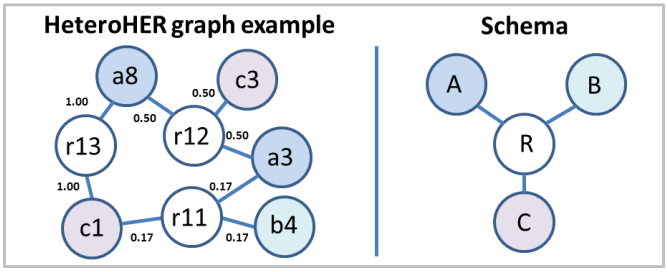

Fig. 3: The graph on the left shows a HeteroHER graph extracted from the example in Fig. 1. For instance, there is a link between r_{11} (the first record of p_1) and a_3 (the third item of category A) if the result of a_3 is abnormal in r_{11}. The link is weighted using Eq. 1. The star-shaped schema on the right is a type-level schema of such a graph.

of risk prediction. A simple function $g(\cdot)$ can be defined as:

$$g(t) = (t - s + 1)/l \tag{1}$$

where t is the time of current record, l is the time window of interest, and s is the starting time of the time window.

Other functions such as truncated Gaussian distribution and Chi Squared distribution can also be used [31]. The window length is the time period of records considered by the model. Note that the window length only sets the scope. It is the link weighting function that controls the contribution of time t records to the model. The two should be considered together according to domain knowledge and/or experimentally.

We include Fig. 3 as an example based on the records of participant p_1 in Fig. 1 to illustrate the process. In this simplified example, we assume all the values of examination items are binary. Different types of examination items in Fig. 1 are treated as different types of nodes on the graph. An abnormal result of the i^{th} item of type Z in the j^{th} record of the k^{th} participant is represented as a link between nodes r_{kj} and z_i. For instance, there is a link between r_{11} and a_3 in the left sub-figure of Fig. 3, and the weight of the link is $(2005 - 2005 + 1)/6 \cong 0.17$ using Eq. (1) with a window width equal to 6 years.

The output of the graph construction process is a heterogeneous graph represented as a set W of sparse matrices W_{ij} for any two node types i, j that are linked to each other in the schema in Fig. 3.

4.1.2 Normalized Weights

To strengthen the weights in the low density region and weaken the weights in the high density region, the weights W_{ij} for $i, j = 1, ..., m$ are further normalized by the row sum and column sum as in [37]:

$$\tilde{W}_{ij} = D_{ij}^{-1/2} W_{ij} D_{ji}^{-1/2} \tag{2}$$

where $d_{ij,pp} = \sum_q W_{ij,pq}$ is the sum of row p in W_{ij} and D_{ij} is an n_i-by-n_i diagonal matrix with the (p, p) element as $d_{ij,pp}$.

4.2 Semi-supervised Learning on HeteroHER Graph

The second component of our method is a semi-supervised learning algorithm for the constructed HeteroHER graph (Section 4.1). The algorithm combines the advantages of [37] for class discovery and [35] for handling heterogeneity to solve a specific problem induced by evidence-based risk prediction from health examination records.

In this section, we first define an objective function for the learning problem and show its convexity, followed by an optimization procedure to solve the problem. Then we derive an efficient iterative algorithm and show its convergence. Finally, time complexity is discussed.

4.2.1 Notations

Let us start with definitions and notations for the following discussions. Assume there are c classes and there is one additional "unknown" class for the cases that are not known to belong to any of the c disease classes. In this work we attach label information of a participant to the Record type nodes representing their examination records. However, the model is general enough to include labels for different types of nodes. Define $Y = [Y_1, ..., Y_m]^{\mathrm{T}} \in \{0, 1\}^{\sum_i^m n_i \times (c+1)}$ such that $Y_i = [y_{i1}, ..., y_{in_i}]^{\mathrm{T}} \in \{0, 1\}^{n_i \times (c+1)}$ encodes the labels of type i nodes. Let $y_{ip}^{(k)}$ be the k^{th} element of vector y_{ip}. If x_{ip}, i.e., node p of type i, is labeled, $y_{ip}^{(k)} = 1$ if x_{ip} belongs to class k; otherwise $y_{ip}^{(k)} = 0$. If x_{ip} is unlabeled, $y_{ip}^{(c+1)} = 1$. By doing so, we set the initial labels of the unlabeled data to be the unknown class. However, we will show later (Section 4.3) that these initial labels for the unlabeled data have little influence on learning their labels.

In addition, we designed the computed labels to be soft labels. Soft labels are especially desirable for medical applications because knowing to what degree of certainty a person is classified into is sometimes as important as knowing the class itself. Let $F = [F_1, ..., F_m]^{\mathrm{T}} \in \mathbb{R}^{\sum_i^m n_i \times (c+1)}$ be the computed soft labels of m node types such that $F_{ip} \in \mathbb{R}^{c+1}$ is a vector indicating the degree of certainty that x_{ip} belongs to any of the $c+1$ classes. The class label of x_{ip} is computed as $\arg\max_{k \le (c+1)} F_{ip}^{(k)}$. F_i can be initialized uniformly amongst type i nodes for $i = 1, ..., m$.

4.2.2 Objective Function

We considered a regularized framework on a heterogeneous graph for our problem. Denote $tr(\cdot)$ as trace and denote $\|\cdot\|_F$ as the Frobenius norm of matrix, i.e., $\|M\|_F^2 = tr(M^T M)$. The classification problem can be viewed as an optimization problem that minimizes an objective function $J(F)$:

$$\begin{aligned} J(F) = &\sum_{ij}^{m} \gamma_{ij} \sum_{p}^{n_i} \sum_{q}^{n_j} \tilde{W}_{ij,pq} \left\| F_{ip} - F_{jq} \right\|_F^2 \\ &+ \sum_{i}^{m} \sum_{p}^{n_i} \mu_{ip} \tilde{d}_{ip} \left\| F_{ip} - Y_{ip} \right\|_F^2 \end{aligned} \tag{3}$$

where $\tilde{W}_{ij}$ is the normalized weights on the links between type i and j nodes as defined in Eq. (2), and F and Y are the same as defined in Section 4.2.1.

The first term is the *smoothness* constraint based on the assumption that the computed labels between the connected nodes in the graph should be close. Let $z = [z_1, ..., z_m]^T$ such that $0 \le z_i \le 1$ be the weights for m node types. Then,

γ_{ij} is defined as between-type weight of type i and type j nodes as follows:

$$\gamma_{ij} = \begin{cases} \frac{1}{2} z_j & \text{if } i = j \\ z_j & \text{otherwise} \end{cases} \quad (4)$$

where the weight is reduced to half for links between same-type nodes to avoid double counting in the summation.

The second term is the *fitness* constraint that penalizes when the computed F is different from labels Y. Let $\tilde{d}_{ij,pp} = \sum_q \tilde{W}_{ij,pq}$ and $\tilde{D}_{ij}$ be an n_i-by-n_i diagonal matrix with the (p,p) element as $\tilde{d}_{ij,pp}$. Then $\tilde{d}_{ip}$ in Eq. (3) is the degree of node p of type i, weighted by its connected node type, i.e., $\tilde{d}_{ip} = \sum_j^m \sum_p^{n_i} z_j \tilde{D}_{ij,pp}$. On the other hand, $\mu_{ip} > 0$ controls the trade-off between the smoothness and fitness constraints for node p of type i.

4.2.3 Convexity

The strict convexity of Eq. (3) minimization is derived in this section. For clarity, we will discuss two terms in Eq. (3) separately. The first term of the objective function denoted as $J_1(F)$ can be derived as follows:

$$\begin{aligned} J_1(F) =& \sum_i^m (\sum_j^m \gamma_{ij} \sum_p^{n_i} \sum_q^{n_j} \tilde{W}_{ij,pq} \left\| F_{ip} - F_{jq} \right\|_F^2) \\ =& \sum_i^m \sum_j^m \gamma_{ij} tr(F_i^T \tilde{D}_{ij} F_i - 2F_i^T \tilde{W}_{ij} F_j + F_j^T \tilde{D}_{ji} F_i) \\ =& tr(\sum_i^m \sum_{i \neq j}^m z_j (F_i^T \tilde{D}_{ij} F_i - 2F_i^T \tilde{W}_{ij} F_j + F_j^T \tilde{D}_{ji} F_i) \\ & + z_i F_i^T (\tilde{D}_{ii} - \tilde{W}_{ii}) F_i) \end{aligned} \quad (5)$$

where $\tilde{W}_{ij}$ is defined in the same way as in Eq. (2) and $\tilde{D}_{ij}$ and γ_{ij} the same as in Section 4.2.2.

Suppose the total number of nodes $n = \sum_i^m n_i$. Let $\tilde{L}$ be a n-by-n block matrix. Let its (i,j) block $\tilde{L}_{ij} = \tilde{D}_{ij} - \tilde{W}_{ij}$ be a Laplacian matrix with normalized weights. Eq. (5) can be transformed to the following matrix expression:

$$J_1(F) = tr(F^T I_z \tilde{L} F) = tr(F^T H F) \quad (6)$$

where I_z is a block diagonal matrix with the elements of the diagonal of (i,i) block equal to z_i, and $H = I_z \tilde{L}$.

The second term of the objective function can be derived as follows:

$$\begin{aligned} J_2(F) &= \sum_i^m \sum_p^{n_i} \mu_{ip} \tilde{d}_{ip} \left\| F_{ip} - Y_{ip} \right\|_F^2 \\ &= tr(\sum_i^m (F_i - Y_i)^T U_i \tilde{D}_i (F_i - Y_i)) \end{aligned} \quad (7)$$

where the diagonal matrix $\tilde{D}_i = \sum_j^m z_j \tilde{D}_{ij}$ and its (p,p) entry is $\tilde{d}_{ip}$ as defined earlier in Section 4.2.2. U_i is a diagonal matrix such that $U_{i,pp} = \mu_{ip}$.

Combining Eq. (6) and Eq. (7), the objective function can be transformed into:

$$J(F) = tr(F^T H F) + tr((F - Y)^T U \tilde{D} (F - Y)) \quad (8)$$

where $\tilde{D}$ is a block diagonal matrix with the diagonal of the (i,i) block equal to $\tilde{D}_i$.

It is easy to verify that $\tilde{L}$ is positive semi-definite and likewise, H, U, $\tilde{D}$, $U\tilde{D}$ and their traces. Therefore the objective function is strictly convex.

4.2.4 Optimization Procedure

The closed-form solution for minimizing Eq. (8) can be obtained by setting the partial derivative of $J(F)$ with respect to F to zero:

$$\left. \frac{\partial J(F)}{\partial F} \right|_{F=F^*} = 2(HF^* + \tilde{D}U(F^* - Y)) = 0 \quad (9)$$

where we use the fact that $H = I_z \tilde{L}$ is symmetrical and that $\tilde{D}$ and U are diagonal.

By multiplying $\tilde{D}^{-1}$ on both sides of Eq. 9 and rearranging the equation, we have:

$$F^* = (\tilde{D}^{-1} I_z \tilde{L} + U)^{-1} U Y \quad (10)$$

By using the fact that $U = \frac{I - I_\alpha}{I_\alpha} = \frac{I_\beta}{I_\alpha}$, we have:

$$\begin{aligned} F^* &= (\tilde{D}^{-1} I_z \tilde{L} I_\alpha + I_\beta)^{-1} I_\beta Y \\ &= (I_\alpha (I - I_z P) + I_\beta)^{-1} I_\beta Y \\ &= (I - I_\alpha I_z P)^{-1} I_\beta Y \\ &= (I - \hat{P})^{-1} I_\beta Y \end{aligned} \quad (11)$$

where $\tilde{L} = \tilde{D} - \tilde{W}$ as before, $P = \tilde{D}^{-1} \tilde{W}$, and $\hat{P} = I_\alpha I_z P$.

Note that the $\infty-$norm of $I_\alpha I_z P$ is lower than 1 given $0 \leq z_i \alpha_i < 1 (i = 1, ..., n)$. Hence the spectral radius of $\hat{P}$ is not greater than the $\infty-$norm. So $(I - \hat{P})$ is invertible.

4.3 Iterative Solution

An iterative algorithm is often more efficient than a closed-form solution with matrix inverse. Here we describe an iterative solution for F and prove its convergence. The optimal F_i for type i nodes can be computed with the following update rule for $i = 1, ..., m$:

$$F_i(t+1) = I_{\alpha_i} \sum_j^m z_j P_{ij} F_j(t) + I_{\beta_i} Y_i \quad (12)$$

where $P_{ij} = \tilde{D}_i \tilde{W}_{ij}$, $I_{\beta_i} = I - I_{\alpha_i}$, and z_j is a type weight scalar as defined earlier in Section 4.2.2.

Eq. (12) bears a label propagation interpretation. Each node iteratively spreads label information to its neighbors until a global stable state is achieved. Particularly, $z_j P_{ij}$ can be seen as the normalized links from type j nodes to type i nodes, scaled by the source type weight z_j. Soft labels of type i nodes at $(t+1)$ are determined by two factors, 1) the computed label scores of neighboring nodes at time t propagated via links, and 2) the initial labels for type i nodes. The diagonal matrix I_{α_i} controls the trade-off between these two influences.

It is important to note that I_{α_i} provides a mechanism to learn an extra outlier class via an instance-level control over the trade-off. Specifically, α_l and α_u are introduced as parameters in the range of $[0, 1]$ that control the influence from the labeled data and unlabeled data respectively. The parameters α_{ip} $(p = 1, ..., n_i)$ in I_{α_i} are defined as $\alpha_{ip} = \alpha_l$ if x_{ip} is labeled, and $\alpha_{ip} = \alpha_u$ if otherwise. The larger α_l and α_u are, the less influence initial labels from Y has.

Algorithm 1 SHG-Health

Input: a set of health examination records of n participants S, the corresponding encoded labels Y
Output: optimized F as the computed soft labels
1: $W \leftarrow$ graph construction from S (Section 4.1.1).
2: Calculate the normalized weights for $i, j = 1, ..., m$ by: $\tilde{W}_{ij} = D_{ij}^{-1/2} W_{ij} D_{ji}^{-1/2}$ (2)
3: Initialize F_i uniformly amongst type i nodes for $i = 1, ..., m$.
4: $t = 1$
5: **repeat**
6: Update F_i for $i = 1, ..., m$ by:
$F_i(t+1) = I_{\alpha_i} \sum_j^m z_j P_{ij} F_j(t) + I_{\beta_i} Y_i$ (12)
7: $t = t + 1$
8: **until** convergence
9: **return** F

Particularly, when α_u is set to a value extremely close to 1, it means that the initial labels of the unlabeled data play almost no role in the learning so that the computed label for an unlabeled case is basically determined by its connectivity in the graph. This mechanism allows the algorithm to learn an additional $(c+1)$ class for nodes that are less connected to the labeled nodes from high risk disease classes.

The complete algorithm of SHG-Health, combining the graph construction and iterative solution, is summarized in Algorithm 1.

4.3.1 Convergence

The proof of the convergence of Eq. (12) is as follows: Let I_α, I_β, z_j and $P = \hat{D}^{-1}\hat{W}$ be the same as defined earlier in Section 4.2.4. The update rule Eq. (12) for type $i = 1, ..., m$ can be reorganized as:

$$\begin{aligned} F_i(t+1) &= I_{\alpha_i} \sum_j^m z_j P_{ij} F_j(t) + I_{\beta_i} Y_i \\ &= I_{\alpha_i} P_i I_{z_i} F(t) + I_{\beta_i} Y_i \\ &= \hat{P}_i F(t) + I_{\beta_i} Y_i \end{aligned} \tag{13}$$

where $\hat{P}_i = I_{z_i} I_{\alpha_i} P_i$. It is equivalent to the following expression:

$$F(t+1) = \hat{P} F(t) + I_\beta Y \tag{14}$$

It has been proved in [37] that $F^* = \lim_{t\to\infty} F(t) = (I - \hat{P})^{-1} I_\beta Y$, which is equivalent to the close-form solution expressed in Eq. (11) and hence our proof is completed.

4.3.2 Time Complexity Analysis

Now we analyze the computational time complexity of the iterative solution (Step 3-8 of Algorithm 1). Step 3 takes $O(k|V|)$ time for initialization, where k is the number of classes (i.e., $k = c + 1$) and $|V|$ is the number of nodes in the graph. At each iteration of Step 6, every link needs to be processed twice, once for the node at each end of the link. This is done for every class, and consequently takes $O(k|E|)$ time, where $|E|$ is the number of links. Also, another $O(k|V|)$ time is needed for incorporating $I_{\beta_i} Y_i$. Therefore, the total time for each iteration is $O(k(|E|+|V|))$, and the total time complexity for the entire iterative solution is $O(lk(|E| + |V|))$, where l is the number of iterations.

5 Experiments

In this section, we evaluate SHG-Health using both real-world datasets and synthetic datasets.

5.1 Datasets

5.1.1 Real Datasets

The real datasets contain a geriatric health examination (GHE) dataset and a Cause of Death (COD) dataset, linked together via the common attribute Person ID, revealing the association between examination results and main cause of death. Records of participants with non-health-related COD were excluded for the purpose of risk identification.

GHE dataset is a de-identified dataset with all private information, such as names, contact details, and birth dates removed. The dataset has 230 attributes, containing 262,424 check-ups of 102,258 participants aged 65 or above, collected during a period of six years (2005 - 2010). The overall ratio of male to female participants is 1.03:1. Each de-identified GHE record is represented by a Person ID and the examination results from a wide range of lab tests, physical examinations, the Brief Symptom Rating Scale (BSRS) mental health assessment, the Short Portable Mental Status Questionnaire (SPMSQ) cognitive function assessment, (de-identified) demographics as well as personal health-related habits, such as exercise, eating, drinking, and smoking habits. Key attributes are listed in Table 1.

COD dataset. The GHE dataset was linked to the Taiwan National Death Registry system using participants' identification numbers and then encrypted to provide de-identified secondary data maintained by the Department of Health of the Taipei City Government. We called this linked subset of data the Cause of Death (COD) dataset. The main causes of death are encoded with the WHO International Classification of Diseases [41], with 9th Revision (ICD-9) in years (2005 - 2008) and ICD-10 in years (2009 - 2010), a standard medical ontology for disease classification. There are in total 522 ICD-9 codes and 925 ICD-10 codes used in the COD dataset. Attributes available from the linked information include a 3-4 digit ICD code for main cause of death and time of death (month and year). For the purpose of risk prediction, we only included those who passed away within three years of their last examination record. This left us with 7,569 (7.4%) participants with COD codes.

The details of GHE attributes and value handling are as follows. For patient demographics, age was firstly discretized into five-year bins, i.e., [65, 70), [70, 75), [75, 80), [80, 85), and [85+. Weights (kilogram) and heights (meter) were used to compute the Body Mass Index (BMI)=weight/(height)^2. BMI is then recorded as an ordinal feature, according to the standard BMI categorization thresholds. For patient habits, most attributes are binary except reason-for-taking-medicine which is a patient reported field. The top 7 most frequently reported reasons were extracted as 7 binary attributes, each indicating the occurrence of a reason. The rest of the reasons were discarded. For lab tests and physical examinations, there are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2561278, IEEE Transactions on Knowledge and Data Engineering

TABLE 1: Selected GHE attributes by categories

Type	Category	Attribute (example)
Patient Profile	Demographics	age, marital status, gender, education level, residential suburb
	Habits	reasons-for-taking-medicine, smoking, drinking, exercise, drink-milk, eat-vegetable, clean-teeth
Lab Tests	Biochemical	glu-ac, total cholesterol (tcho), thyroglobulin (tg), got, gpt, albumin (alb), thyroid stimulating hormone (tsh)
	Blood	red blood cell, white blood cell, plate, hematocrit (hct), mean corpuscular volume (mcv), mean corpuscular hemoglobin (mch), alpha-fetoprotein (afp), hemoglobin (hb)
	Urine	outlook, ph, protein, sugar, blood, red blood cell, white blood cell, pus cell, epithelium cell, casts
	Other	faecal occult blood test (fobt)
Examinations	Physical	weight, height, waist, systolic blood pressure, diastolic blood pressure, pulse rate
	External	neck, chest, heart, breast, abdomen, back, rectum, limbs, prostate
	Other	X-ray, EKG, cervical smear, abdominal ultrasound
Mental Health	BSRS	5 questions regarding nervousness, anger, depression, comparison with others, and sleep
Cognitive Function	SPMSQ	10 questions, e.g., current date, day of the week, where the person is situated, home address, age, year of birth, etc.

three fields to record the results, namely observed value, status, and description. The observed values are generally numeric. The status fields indicate whether or not the result of a test is normal. Their values can be either binary or ordinal, depending on the type of tests. The descriptions are in free text format. We only used the information from the status fields for the following reasons. Firstly, the reference ranges of these items may differ amongst hospitals and the information regarding where an examination was taken is not available in the dataset for privacy reasons. Secondly, the values for the description fields are mostly missing.

The results of 5 mental health assessment (BSRS) questions were encoded in terms of degree of severity, i.e., normal, mild, medium, and severe. The overall result of the cognitive function assessment (SPMSQ) was scored in terms of the number of questions that were incorrectly answered. It is an ordinal attribute with values "sound" (0-5 scores), "mild" (6-9 scores), "medium" (10-14 scores), and "severe" (above 15 scores) according to the standard categorization. All the resulting discrete attributes discussed above were then binarized if they were not already in the binary form.

The data was collected in a standard annual health examination program for elderly people, run by the Taipei City Government. Participants voluntarily took part in the program, and were encouraged to visit on a yearly basis. Data related to individual identification was removed before the dataset acquisition. The acquisition and processing of the data was approved by the Institutional Review Board (IRB) of the Taipei City Hospital.

5.1.2 Synthetic Datasets

To test the stability of our method, we also generated two groups of synthetic datasets based on the distribution of the processed real datasets. Specifically, we computed the value distributions for a given disease class, a given year, and a given feature for 10 disease classes selected based on COD codes. The selection of the disease classes is reported in Section 5.2.1. The cumulative distribution functions (CDFs) were computed, based on which a random number generator was used to select a value for a given disease class, a year, and a feature. The resulting synthetic datasets are:

- **Balanced datasets**: to test the stability with increasing class size in a balanced-class setting, we generated same number of instances for all disease classes and the "unknown" (unlabeled) class with increasing class size in the range of $\{100, 300, 500, 1000\}$, and
- **Increasing unlabeled size datasets**: to test the stability given increasing scale of unlabeled data, we generated 1000 instances for every disease class with the "unknown" (unlabeled) class size equals to $1000 \times p$, where p is a scalar in the range of $\{1, 3, 5, 10, 15, 20\}$.

Note that the synthetic datasets were generated based on the assumption that features are independent, which does not hold in real-world healthcare data. However, by using synthetic datasets we can better understand algorithm behaviors in different class settings.

5.2 Experimental Settings

5.2.1 High-risk Disease Classes

We selected 10 ICD disease categories as high-risk disease classes. The first 3 digits of ICD 10 were used to define disease categories and mapped to the corresponding 3-digit ICD 9 codes for the records taken before 2009. Table 2, below, shows the size of the 10 disease classes in terms of the number of participants and number of records.

The number of classes was determined with the intention to maximize the number of diseases to show a wide coverage of disease categories, given the constraint on the number of disease instances available. To make sure the number of instances per class is sufficient for reasonable classification, we selected diseases with top-10 frequency counts. Based on the suggestions from clinical experts, we excluded diabetes mellitus, which is known to have many complications, and acute myocardial infarction, an acute disease. This gave us the first 8 classes in Table 2. To make the problem closer to a real-world situation, we selected two additional diseases with less frequency to represent

TABLE 2: Sizes of 10 disease categories and unlabeled cases in terms of participant (P) and record (R).

Class	ICD10	Name	ICD9	Size (P)	Size (R)
1	J18	Pneumonia, organism unspecified	481,485,486,514	434	851
2	J40-J44	Chronic lower respiratory diseases	490-493,496	277	531
3	C34	Malignant neoplasm of bronchus and lung	162	253	586
4	C22	Malignant neoplasm of liver and intrahepatic bile ducts	155	219	375
5	C16	Malignant neoplasm of Stomach	151	193	355
6	C18	Malignant neoplasm of colon	153	186	371
7	I25	Chronic ischaemic heart disease	412,414,429	185	330
8	C25	Malignant neoplasm of pancreas	157	123	242
9	C23	Malignant neoplasm of gallbladder	156	74	126
10	G20	Parkinson's disease	332	37	73
Unlabeled (alive) cases				26,771	69,802

minority classes. The unlabeled cases were then randomly selected from the alive cases with a ratio equal to data distribution i.e., the deceased/alive ratio as 7.4/92.6. This gave us 69,802 records from 26,771 unlabeled participants (Table 2, above).

5.2.2 HeteroHER Graph Construction

Four node types were modeled for constructing our HeteroHER graph, namely *Record*, *Physical Test*, *Mental Assessment*, and *Profile*. *Physical Test* refers to all the lab tests and physical examinations in Table 1, while *Mental Assessment* covers both the mental health assessments and cognitive function assessments. *Profile* includes all the patient demographics and habits, and *Record* indicates the artificial nodes created for representing individual records. As discussed earlier, only the abnormal results, both from physical and mental tests, were included in graph construction. Table 3 shows the network statistics of the HeteroHER graphs extracted from both the real and synthetic datasets. The numbers of nodes for *Physical Test* (Test), *Mental Assessment* (Mental), and *Profile* types refer to the total number of distinct values of the extracted features for the type. Note that density is calculated as the ratio of the number of edges E to the number of possible edges P. Since HeteroHER Graph has a star schema (Fig. 3), P is calculated as the number of *Record* nodes times the number of all attribute type nodes.

5.2.3 Evaluation Metrics

We designed a two-stage evaluation strategy to evaluate the proposed SHG-Health.

In the first stage, we evaluated an algorithm's ability to identify high-risk cases, regardless of their disease category. It can be understood as testing the algorithm's ability to predict mortality risk. All predicted disease cases were regarded as *predicted positive* cases and the true disease cases were regarded as (true) *positive* cases. As there is no ground truth for the negative or healthy cases, we used measures that focus on positive predictions, namely precision, recall/sensitivity, and F-score. While precision measures how correct the positive predictions of an algorithm are, recall shows its ability to catch the positive cases. F-score calculates the harmonic mean between precision and recall.

In the second stage, we looked into a method's ability to predict the correct disease class given that it predicted a case to be in one of the high-risk classes. This is a conditional evaluation that only considers cases that were predicted as one of the disease classes. Macro-precision and macro-recall measures were used. Macro-averaging takes the average of precision or recall scores computed from individual classes [42]. It assumes that all classes are equally important, so that the performance of minority classes can be reflected in the macro-averaged scores.

5.2.4 Algorithms for Comparison

For SHG-Health, we compared three time-weighted functions for graph construction, namely Eq. (1) (Ours), truncated Gaussian (Ours-Gaus), and truncated Chi Squared (Ours-Chi2) discussed in Section 4.1.1. Other algorithms that were compared with our method are:

- **Support Vector Machines:** SVM has been adopted as one of our baseline methods. Although SVM with RBF kernel achieved the best results in [8], linear and RBF kernels had very similar performance in our experiments. We only report the results of the linear kernel for its favourable efficiency. The LIBSVM [43] and LIBLINEAR [44] implementations were used in our experiments for the RBF and linear kernels respectively.
- **Nearest Neighbor Classifier**: KNN classifier is a common baseline for graph-based models. K was experimentally set to 1.
- **General Graph-based Semi-Supervised Learning:** GGSSL [37] is a state-of-the-art graph-based semi-supervised method for class discovery. As it is not directly applicable to heterogeneous graphs, we constructed a homogeneous graph by converting all types of nodes in our heterogeneous graph into a single-type graph. The MATLAB implementation available from the author webpage [45] was employed in our experiments.
- **GNetMine:** GNetMine [35] is a state-of-the-art graph-based semi-supervised method on a graph of heterogeneous nodes. The MATLAB implementations by the authors, available from GitHub [46], was employed in our experiments.

For parameter tuning, the parameters c and γ for SVM were tuned based on the $\{10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$ grid. The parameters α_i and λ_{ij} in GNetMine denote type weights and type relationship weights respectively. They were tuned based on the α_i/λ_{ij} ratio grid $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ with α_i fixed to 0.1 as in [35]. For fair comparison to GNetMine, the type weight parameters z_j $(j = 1, ..., m)$ of SHG-Health were also set uniformly. The parameters

TABLE 3: Extracted HeteroHER graph statistics

Dataset		# people	# nodes					# links to Record nodes				Density
			Record	Test	Mental	Profile	Total	Test	Mental	Profile	Total	
Real	GHE@10class	26,771	73,642	55	26	55	73,778	601,062	119,952	523,387	1,244,401	0.1242
Synthetic	(100,100)	1,100	3,013	55	26	55	3,149	28,071	4,611	22,552	55,234	0.1348
	(300,300)	3,300	9,054				9,190	84,052	13,982	68,053	166,087	0.1349
	(500,500)	5,500	15,092				15,228	139,736	23,134	113,231	276,101	0.1345
	(1000,1000)	11,000	30,201				30,337	280,412	46,655	227,932	554,999	0.1351
	(1000,3000)	13,000	35,463				35,599	323,101	55,263	265,067	643,431	0.1334
	(1000,5000)	15,000	40,695				40,831	365,005	63,974	301,661	730,640	0.1320
	(1000,10000)	20,000	53,674				53,810	469,575	85,167	393,486	948,228	0.1299
	(1000,15000)	25,000	66,841				66,977	575,360	106,694	485,914	1,167,968	0.1285
	(1000,20000)	30,000	79,979				80,115	682,022	128,357	579,269	1,389,648	0.1278

TABLE 4: Evaluation on binary prediction (avg±std%)

	Precision	Recall/Sensitivity	F-Score
Ours	96.24 ± 1.60	**43.93** ± 1.11	**60.32** ± 1.23
Ours-Chi2	**99.33** ± 0.36	43.02 ± 1.40	60.02 ± 1.41
Ours-Gaus	96.99 ± 1.32	43.69 ± 1.14	60.23 ± 0.86
SVM	89.00 ± 10.19	0.49 ± 0.36	0.98 ± 0.71
KNN	37.52 ± 1.48	25.62 ± 1.30	30.45 ± 1.36
GNetMine	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GGSSL	5.21 ± 0.02	100.00 ± 0.00	9.91 ± 0.03

TABLE 5: Evaluation on disease class prediction (avg±std%)

	Macro-Precision	Macro-Recall
Ours	89.14 ± 0.56	89.62 ± 0.38
Ours-Chi2	**90.58** ± 0.19	**90.73** ± 0.15
Ours-Gaus	89.55 ± 0.56	90.30 ± 0.41
KNN	21.12 ± 1.49	59.92 ± 2.50
SVM	52.50 ± 39.41	63.33 ± 30.55
GNetMine	-	-
GGSSL	0.11 ± 0	9.09 ± 0

α_l and α_u that control the influence from Y for the labeled and the unlabeled data respectively, were set to 0.01 for α_l as in [37] and tuned based on the grid $\{0.9, 0.99, 0.999, 0.9999, 0.99999\}$ for both SHG-Health and GGSSL methods.

All of the experiments were conducted using 5-fold stratified cross validation. The performance on the testing set was obtained by averaging over 5-fold results. All the experiments were run on an Intel(R) Core(TM) CPU@3.40GHz workstation with 16GB physical memory.

5.3 Result Analysis

In this section we report and analyze the experimental results. The best $\lambda_{ij}, \forall i, j \in \{1, ..., m\}$ for GNetMine was 0.2 in all of our experiments, the same as reported in [35]. The best α_u for SHG-Health is 0.99, while GGSSL tends to bias toward one disease class with $\alpha_u = 0.9999$ and completely toward the unknown class with α_u less than that.

5.3.1 Identifying High-risk Cases

In this first stage of the two-stage evaluation (Section 5.2.3), we compared algorithms based on their abilities to identify high-risk cases regardless of what disease category they belonged to. All of the cases from different high-risk disease classes were regarded as belonging to one class, i.e., the high-risk class. This binary setting evaluates how well an algorithm is able to pick up high-risk cases in general. Table 4, above, shows that our algorithms achieved the best overall performance at 99.33% precision (Ours-Chi2), 43.93% recall (Ours), and 60.32% F score (Ours). GGSSL had 100.00% recall but extremely low precision at 5.21%, which indicates that it leaned towards predicting most cases as high-risk. On the other hand, GNetMine was completely biased toward the "unknown" class and thus had zero precision, recall, and F scores. The fact that recall scores for all methods are less than 50% also shows that capturing positive cases from large and noisy unlabeled cases is difficult. Our proposed SHG-Health could be seen as a conservative model, which is desirable for a preventive care system because the cost of false alarms is high [14], [47].

5.3.2 Classifying into Correct Disease Categories

In the second stage, we further evaluated the algorithms' conditional performance on multi-class classification. Only the cases that were predicted into one of the disease classes were considered. The macro-averaging measures were used to evaluate how correct these predictions were at the disease category level. Note that any "unknown" case incorrectly predicted as one of the disease classes was counted as an incorrect prediction in this calculation.

Table 5, above, shows that our SHG-Health, especially Ours-Chi2, outperformed all the other algorithms, achieving 90.58% macro-precision and 90.73% macro-recall. Overall, our algorithm is able to classify high-risk individuals into a correct disease category quite accurately.

5.3.3 Top Scored Test Items

Based on the COD labels available for the record type nodes, SHG-Health also computes scores as soft labels for other types of nodes, such as the *Physical Test* nodes. Within-class scores of a node type can reveal the relative importance of those nodes in the class.

Top-5 scored *Physical Test* items for 10 disease classes are listed in Table 6, below. There are some interesting results identified by our clinical experts. For example, for lung related disease categories, namely *pneumonia*, *chronic lower respiratory diseases*, and *malignant neoplasm of bronchus and lung*, chest examination has the highest score. This bears the interpretation that participants with these diseases commonly have abnormal chest examination results. Another obvious example can be found in the *malignant neoplasm of liver and intrahepatic bile ducts* class. Top-1 ranked item Alpha-Fetoprotein is a commonly used tumor marker for

TABLE 6: Top 5 scored test items for 10 disease classes

Top 5	1) Pneumonia, organism unspecified	2) Chronic lower respiratory diseases	3) Malignant neoplasm of bronchus and lung	4) Malignant neoplasm of liver and intrahepatic bile ducts
1	Chest Exam	Chest Exam	Chest Exam	Alpha-Fetoprotein
2	Albumin	Urinary casts	Alpha-Fetoprotein	Aspartate Aminotransferase (GOT)
3	Urinary casts	Albumin	Hemoglobin	Glutamic-Pyruvic Transaminase (GPT)
4	Hemoglobin	Hemoglobin	Hematocrit	Platelet count
5	Blood urea nitrogen (BUN)	Alpha-Fetoprotein	Mean corpuscular volume	Glucose urine test

Top 5	5) Malignant neoplasm of Stomach	6) Malignant neoplasm of colon	7) Chronic ischaemic heart disease	8) Malignant neoplasm of pancreas
1	Albumin	Albumin	Creatinine blood test	Glucose urine test
2	Hemoglobin	Alpha-Fetoprotein	Blood urea nitrogen (BUN)	Albumin
3	Mean corpuscular volume	Chest	Glucose urine test	Mean corpuscular volume
4	Platelet count	Hemoglobin	Protein in Urine (Proteinuria)	Alpha-Fetoprotein
5	Hematocrit	Creatinine blood test	Hemoglobin	Pus Cell in Urine

Top 5	9) Malignant neoplasm of gallbladder	10) Parkinson's disease
1	Alpha-Fetoprotein	Urinary casts
2	Aspartate Aminotransferase (GOT)	Chest
3	Chest	Hemoglobin
4	Mean corpuscular volume	Albumin
5	Red blood cell count	Red blood cell count

liver cancer. GOT and GPT are the enzymes concentrated in the liver, commonly used as key indicators for evaluating liver damage.

These results show that SHG-Health is able to identify important examination items for disease classes. By modeling features (i.e., examination items) as different types of nodes on a graph, the computing of soft labels for these nodes is actually a mechanism of feature weighting. It is the connection to these highly scored features of a class that determines the class label of a *Record* node in the graph.

5.3.4 Stability on Synthetic Data

We further compared the stability of the algorithms using two groups of synthetic data generated based on the distribution of the real data. The first group is the balanced datasets with increasing size per class in the range of $\{100, 300, 500, 1000\}$. The second group contains the datasets with increasing number of unlabeled cases in the scale of $\{1, 3, 5, 10, 15, 20\}$ times of the size of a labeled class. For the details on how these datasets were generated, please refer to Section 5.1.2. We evaluated algorithms in terms of their ability to identify high-risk cases (Task 1) and their ability to classify a high-risk case into the correct disease class (Task 2). The same measures as in the real dataset case were used. Due to space limitations, it suffices to report the F scores for Taks 1 and macro-averaging scores for Task 2.

Figure 4, below, shows the results on the balanced synthetic datasets. It can be seen from the F scores that all the algorithms performed stably in Task 1 except that 1NN had lower scores when class sizes were small. Our method is comparable to SVM and they had the highest performance in Task 1. However, our algorithm achieved significantly better macro-precision and macro-recall scores than the other algorithms in Task 2. While our SHG-Health stably maintained 70% macro-precision, others fluctuated below 50%. A similar phenomenon can be observed in the case of macro-recall.

We expected the performance would drop with the increasing scales of unlabeled data. The F scores of Task 1 confirm this intuition in Figure 5. The 1NN, GNetMine, and GGSSL methods had a steeper descending gradient than SVM and our approach, as the size of the unlabeled cases increases from 1,000 to 20,000. However, when it comes to predicting correct disease classes (Task 2), our SHG-Health had the highest macro-precision and macro-recall scores on the synthetic datasets. Note that the performance of our method went up as the unlabeled sizes increased from 10,000 to 20,000. It could be that more unlabeled cases helped our method to differentiate between disease classes better. It is worth mentioning that SVM had the most stable performance in Task 2, slightly below 50% for both macro-averaging measures.

5.3.5 Time Analysis

To investigate the scalability of all the algorithms, we recorded the training time for experiments on the synthetic datasets (Section 5.3.4), except that the testing time of 1NN was recorded. Figure 6, below, compares the time performance of all the algorithms with increasing data sizes. It can be seen that our method is the most time-efficient method of all, with only 1.41 seconds of training time at $(1000, 20000)$, i.e., 10×1000=10,000 disease cases plus 20,000 unlabeled cases. Note that the implementations we used for other methods are either standard implementations (SVM and 1NN) or from the author provided codes (GNetMine and GGSSL) and they might not be best tuned for efficiency. The purpose of Figure 6 is to show that our approach also enjoys desirable efficiency given its superior effectiveness as demonstrated in previous sections.

5.3.6 Discussion

The experimental results showed that SHG-Health performed the best amongst all the algorithms compared. Particularly, its ability 1) to identify high-risk cases and 2) to predict the correct disease category for high-risk cases has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2561278, IEEE Transactions on Knowledge and Data Engineering

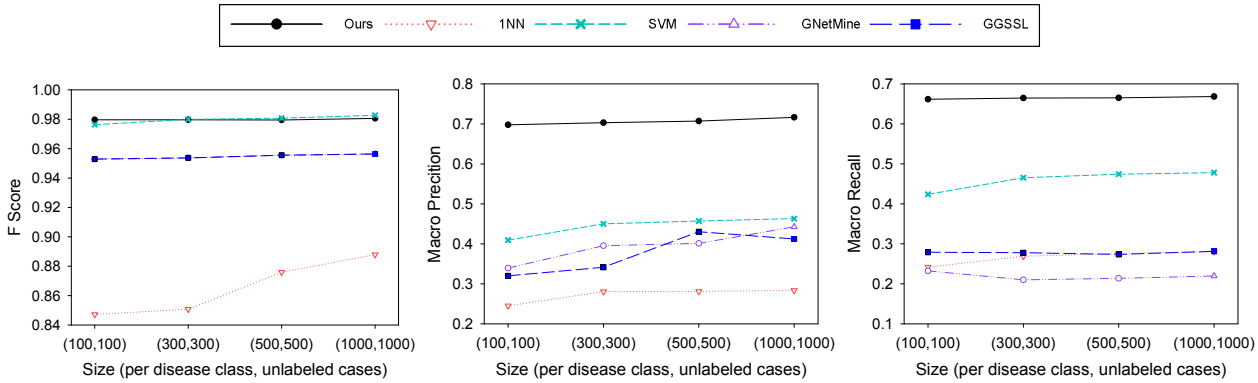

Fig. 4: Results of the balanced synthetic datasets with increasing class sizes.

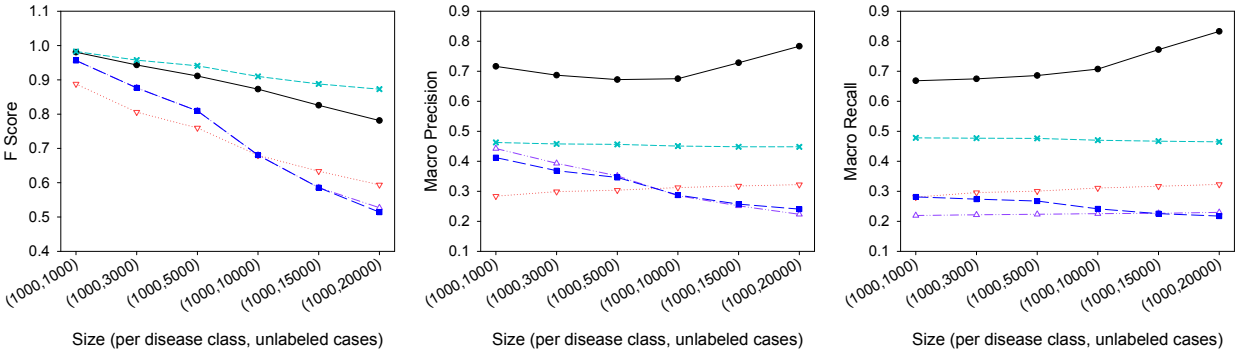

Fig. 5: Results of the synthetic datasets with increasing unlabeled cases.

been demonstrated and verified on our real datasets and synthetic datasets (Section 5.1).

From the results of real datasets reported in Table 4 and Table 5, we can see that although GNetMine also utilizes a heterogeneous network structure for classification, it tends to be biased toward the noisy "unknown" class. The reason could be that GNetMine does not have a mechanism to control the label influence at the instance level, such as the α parameters (Section 4.3) in our method and in GGSSL, nor the ability for class discovery. On the other hand, although GGSSL has such a mechanism, a homogeneous graph construction missed out the type-specific information that can help the classification. As a result, GGSSL was completely biased toward the dominant disease class.

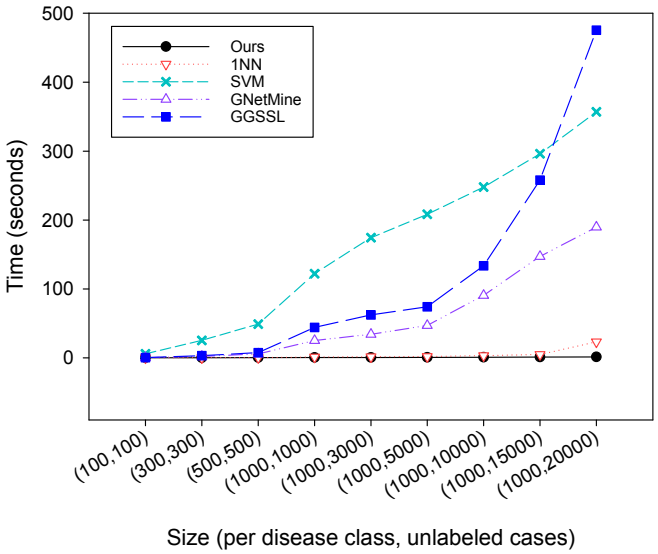

Fig. 6: Computational time analysis on the synthetic datasets for the algorithms compared.

In the case of synthetic datasets, the performance for most methods dropped in Task 2 (Fig. 4 and Fig. 5). This can be explained by the information lost due to the feature independence assumption for generating the synthetic datasets. The exceptions are GNetMine and GGSSL, which showed more reasonable performance on synthetic datasets. It could be that they are better in handling independent features than correlated ones. The efficiency of our method has also been demonstrated in Fig. 6 based on the same synthetic datasets.

It is worth noting that the choice of disease combination can affect the performance. For example, when we selected 5 diseases from the ICD-10 "Malignant neoplasms, digestive organs" category in a separate experiment, the disease-level classification performance dropped to below 23% for all algorithms, due to the less discriminability amongst the instances of these diseases.

Overall, although SHG-Health is conservative in predicting cases into high-risk classes, we have shown that it is able to predict the correct disease classes with high scores in all evaluation measures. This is very desirable for the preventive type of Clinical Decision Support Systems (CDSSs). False positives are especially costly in preventive care, which could result in unnecessary anxiety, worry, and invasive diagnostic tests [14], [47]. In addition, it is believed that CDSSs are to *support* clinical professionals rather than to *replace* them. Therefore, a good system should be able to identify and draw attention to participants with high risks.

6 Conclusion

Mining health examination data is challenging especially due to its heterogeneity, intrinsic noise, and particularly the large volume of unlabeled data. In this paper, we introduced an effective and efficient graph-based semi-supervised algorithm namely SHG-Health to meet these challenges.

Our proposed graph-based classification approach on mining health examination records has a few significant advantages.

- Firstly, health examination records are represented as a graph that associates all relevant cases together. This is especially useful for modeling abnormal results that are often sparse.
- Secondly, multi-typed relationships of data items can be captured and naturally mapped into a heterogeneous graph. Particularly, the health examination items are represented as different types of nodes on a graph, which enables our method to exploit the underlying heterogeneous subgraph structures of individual classes to achieve higher performance.
- Thirdly, features can be weighted in their own type through a label propagation process on a heterogeneous graph. These in-class weighted features then contribute to the effective classification in an iterative convergence process.

Our work shows a new way of predicting risks for participants based on their annual health examinations. Our future work will focus on the data fusion for the health examination records to be integrated with other types of datasets such as the hospital-based electronic health records and the participants' living conditions (e.g., diets and general exercises). By integrating data from multiple available information sources, more effective prediction may be achieved.

Acknowledgments

This study is based on data from the Taipei City Public Health Database provided by the Department of Health, Taipei City Government, and managed by Databank for Public Health Analysis (DoPHA). The interpretation and conclusions contained herein do not represent those of Department of Health, or DoPHA. The research is partially funded by the Australian Research Council Discovery Project ID DP140100104.

References

[1] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostics," *IEEE International Conference on Data Mining*, pp. 201–210, 2013.

[2] T. Tran, D. Phung, W. Luo, and S. Venkatesh, "Stabilized sparse ordinal regression for medical risk stratification," *Knowledge and Information Systems*, pp. 1–28, Mar. 2014.

[3] M. S. Mohktar, S. J. Redmond, N. C. Antoniades, P. D. Rochford, J. J. Pretto, J. Basilakis, N. H. Lovell, and C. F. McDonald, "Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data," *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 51–59, 2015.

[4] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 381–391, 2010.

[5] E. Kontio, A. Airola, T. Pahikkala, H. Lundgren-Laine, K. Junttila, H. Korvenranta, T. Salakoski, and S. Salanterä, "Predicting patient acuity from electronic patient records." *Journal of Biomedical Informatics*, vol. 51, pp. 8–13, 2014.

[6] Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information." *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, no. 3, pp. 501–8, 2014.

[7] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," *IEEE Transactions Knowledge and Data Engineering*, vol. 27, no. 1, pp. 130–141, 2015.

[8] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in *2014 IEEE International Conference on Data Mining*, 2014, pp. 761–766.

[9] S. Pan, J. Wu, and X. Zhu, "CogBoost: Boosting for Fast Cost-sensitive Graph Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, no. 1, pp. 1–1, 2015.

[10] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "A survey and analysis of Electronic Healthcare Record standards," *ACM Computing Surveys*, vol. 37, no. 4, pp. 277–315, 2005.

[11] C. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li, "Cognitive impairment assessed at annual geriatric health examinations predicts mortality among the elderly," *Preventive Medicine*, vol. 67, pp. 28–34, 2014.

[12] "Health assessment for people aged 75 years and older," http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare_mbsitem_75andolder, accessed: 2015-05-03.

[13] "Health checks for the over-65s," http://www.nhs.uk/Livewell/Screening/Pages/Checksover65s.aspx, accessed: 2015-05-03.

[14] L. Krogsbøll, K. Jørgensen, C. Grønhøj Larsen, and P. Gøtzsche, "General health checks in adults for reducing morbidity and mortality from disease (Review)," *Cochrane Database of Systematic Reviews*, no. 10, 2012.

[15] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015.

[16] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, no. 4, pp. 613–618, 2013.

[17] H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2314–2320, 2012.

[18] T. P. Nguyen and T. B. Ho, "Detecting disease genes based on semi-supervised learning and protein-protein interaction networks," *Artificial Intelligence in Medicine*, vol. 54, no. 1, pp. 63–71, 2012.

[19] V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 869–875, 2013.

[20] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu, "Exploring patient risk groups with incomplete knowledge," *IEEE International Conference on Data Mining*, pp. 1223–1228, 2013.

[21] T. Hwang and R. Kuang, "A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery," *SIAM International Conference on Data Mining*, pp. 583–594, 2010.

[22] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal Phenotyping from Longitudinal Electronic Health Records," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pp. 705–714, 2015.

[23] F. Zhang, Y. Song, and W. Cai, "A ranking-based lung nodule image classification method using unlabeled image knowledge," *IEEE International Symposium on Biomedical Imaging*, pp. 1356–1359, 2014.

[24] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," *International Conference on Machine Learning*, vol. 20, no. 2, pp. 912–919, 2003.

[25] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," in *International Conference on Machine Learning*, 2002, pp. 387–394.

[26] P. Yang, X. Li, H. N. Chua, C. K. Kwoh, and S. K. Ng, "Ensemble positive unlabeled learning for disease gene identification," *PLoS ONE*, vol. 9, no. 5, 2014.

[27] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng, "Positive-unlabeled learning for disease gene identification," *Bioinformatics*, vol. 28, no. 20, pp. 2640–2647, 2012.

[28] L. T. Nguyen, M. Zeng, P. Tague, and J. Zhang, "I Did Not Smoke 100 Cigarettes Today! Avoiding False Positives in Real-World Activity Recognition," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2015, pp. 1053–1063.
[29] Y. Zhao, G. Wang, X. Zhang, J. X. Yu, and Z. Wang, "Learning phenotype structure using sequence model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 667–681, 2014.
[30] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8852–8858, Aug. 2012.
[31] L. Chen, X. Li, Y. Yang, H. Kurniawati, Q. Z. Sheng, H.-Y. Hu, and N. Huang, "Personal health indexing based on medical examinations: A data mining approach," *Decision Support Systems*, vol. 81, pp. 54 – 65, 2016.
[32] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch, "Learning with Local and Global Consistency," *Advances in Neural Information Processing Systems*, vol. 1, pp. 595–602, 2003.
[33] A. Guillory and J. Bilmes, "Label Selection on Graphs," in *Neural Information Processing Systems*, 2009, pp. 1–9.
[34] 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*. ACM, 2009, pp. 797–806.
[35] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, "Graph regularized transductive classification on heterogeneous information networks," *Machine Learning and Knowledge Discovery in Databases*, vol. 6321 LNAI, pp. 570–586, 2010.
[36] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1298–1306.
[37] F. Nie, S. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing and Applications*, vol. 19, pp. 549–555, 2010.
[38] M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang, "Compact Graph based Semi-Supervised Learning for Medical Diagnosis in Alzheimer's Disease," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1192–1196, 2014.
[39] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease." *Nature reviews. Genetics*, vol. 12, no. 1, pp. 56–68, 2011.
[40] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.
[41] "International classification of diseases," http://www.who.int/classifications/icd/en/, accessed: 2015-05-05.
[42] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.
[43] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[44] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
[45] F. Nie, "GGSSL code," https://sites.google.com/site/feipingnie/publications, 2010.
[46] M. Ji, "GNetMine," https://github.com/rackingroll/HetePathMine/blob/master/GNetMine.m, 2010.
[47] Z. Jakab, E. Comparative, and T. Eu, "Periodic health examination: A brief history and critical assessment," *Eurohealth*, vol. 15, no. 4, pp. 16–20, 2009.

Ling Chen received her B.A. degree in philosophy from the Department of Philosophy, National Taiwan University, Taiwan in 2006, and the M.S. degree in computer science from the School of Information Technology and Electrical Engineering at The University of Queensland, Australia in 2011. She is currently working towards a Ph.D. degree in computer science at The University of Queensland. Her research interests include data mining, machine learning and their applications in healthcare.

Xue Li received his M.S. and Ph.D. degrees from The University of Queensland and Queensland University of Technology in 1990 and 1997 respectively. Currently, he is a Professor in the School of Information Technology and Electrical Engineering at The University of Queensland in Brisbane, Queensland, Australia. His major areas of research interests and expertise include: Data Mining, Multimedia Data Security, Database Systems, and Intelligent Web Information Systems. He is a member of ACM, IEEE, and SIGKDD.

Quan Z. Sheng received his Ph.D. degree in computer science from the University of New South Wales, Sydney, Australia in 2006. He is a Professor and Deputy Head of the School of Computer Science at the University of Adelaide. His research interests include Web of Things, big data analytics, distributed computing, and Internet technologies. He is the recipient of the ARC Future Fellowship (2014), Chris Wallace Award for Outstanding Research Contribution (2012), Microsoft Research Fellowship (2003) and CSC Fellowship (1998). He has more than 220 publications and is a member of the IEEE and the ACM.

Wen-Chih Peng received his B.S. and M.S. degrees from the National Chiao Tung University, Taiwan, in 1995 and 1997, respectively, and the Ph.D. degree in electrical engineering from the National Taiwan University, Taiwan, R.O.C, in 2001. Currently, he is a Professor in the Department of Computer Science, National Chiao Tung University, Taiwan. Prior to joining the Department of Computer Science, National Chiao Tung University, he was mainly involved in the projects related to mobile computing, data broadcasting, and network data management. His research interests include mobile data management and data mining. He is a member of the IEEE.

John Bennett is a General Practitioner at the University Health Service of The University of Queensland. He is a Fellow of the Royal Australian College of General Practitioners. His medical interests include medical issues of young adults, travel medicine, medical problems of older patients, preventative healthcare, and lifestyle assessments. He holds a Ph.D. in population health and a honours degree in computer science from The University of Queensland, Australia.

Hsiao-Yun Hu is an Assistant Investigator in Department of Education and Research, Taipei City Hospital. She holds a Ph.D. in Public Health from the National Yang-Ming University. Her research interest includes obesity Epidemiology and medical utilization of chronic diseases.

Nicole Huang is a Professor in Institute of Hospital and Health Care Administration at National Yang-Ming University. She holds a Ph.D. in Health Policy and Management from the Johns Hopkins Bloomberg School of Public Health. Her research interest includes Health Policy and Management, Health Services Research, and Health Disparities.