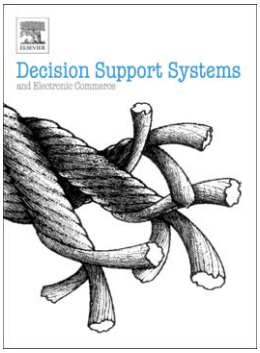


Personal health indexing based on medical examinations: a data mining approach

Ling Chen, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu, Nicole Huang

PII: S0167-9236(15)00202-X
DOI: doi: 10.1016/j.dss.2015.10.008
Reference: DECSUP 12661

To appear in: *Decision Support Systems*

Received date: 5 February 2015
Revised date: 23 June 2015
Accepted date: 29 October 2015

Please cite this article as: Ling Chen, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu, Nicole Huang, Personal health indexing based on medical examinations: a data mining approach, *Decision Support Systems* (2015), doi: 10.1016/j.dss.2015.10.008

Personal health indexing based on medical examinations: a data mining approach

Ling Chen[a,*], Xue Li[a], Yi Yang[a], Hanna Kurniawati[a], Quan Z. Sheng[b], Hsiao-Yun Hu[c,e], Nicole Huang[d,c]

[a]*School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia*
[b]*School of Computer Science, The University of Adelaide, Adelaide, Australia*
[c]*Department of Education and Research, Taipei City Hospital, Taipei, Taiwan*
[d]*Institute of Hospital and Health Care Administration, National Yang-Ming University, Taipei, Taiwan*
[e]*Institute of Public Health and Department of Public Health, National Yang-Ming University, Taipei, Taiwan*

Abstract

We design a method called MyPHI that predicts personal health index (PHI), a new evidence-based health indicator to explore the underlying patterns of a large collection of geriatric medical examination (GME) records using data mining techniques. We define PHI as a vector of scores, each reflecting the health risk in a particular disease category. The PHI prediction is formulated as an optimization problem that finds the optimal soft labels as health scores based on medical records that are infrequent, incomplete, and sparse. Our method is compared with classification models commonly used in medical applications. The experimental evaluation has demonstrated the effectiveness of our method based on a real-world GME data set collected from 102,258 participants.

Keywords: personal health index, geriatric medical examination, label uncertainty, data mining, feature extraction

*Corresponding author
Email addresses: l.chen5@uq.edu.au (Ling Chen), xueli@itee.uq.edu.au (Xue Li), yi.yang@uq.edu.au (Yi Yang), hannakur@uq.edu.au (Hanna Kurniawati), qsheng@cs.adelaide.edu.au (Quan Z. Sheng), A3547@tpech.gov.tw (Hsiao-Yun Hu), syhuang@ym.edu.tw (Nicole Huang)

1. Introduction

Modern societies have experienced dramatic growth in elderly population from the beginning of this century. This implies increasing healthcare needs and government expenditure. For example, the U.S. government spent $414.3 billion in elderly health care in 2011, $100 billion higher than the inflation-adjusted expenses in 2001 [1]. Annual geriatric medical examination (GME) is now an integral part of elderly healthcare for many developed countries. For instance, Australia [2], United Kingdom [3], and Taiwan [4] have GME programs to periodically monitor health status of senior residents. However, it is always a difficult task for healthcare professionals to provide an overall report on personal health after a comprehensive medical check-up is performed with hundreds of parameters. Moreover, the richness of GME records, such as correlations amongst test results, their longitudinal progression, and their relationships to other participants that have similar patterns of health development, is often left unexplored. In fact, such exploration is manually impossible, because the complexity of the combined effects grows exponentially with the growth of the number of different test results, the available number of longitudinal records, and the total number of participants.

We design a method called MyPHI that predicts personal health index (PHI), a new health indicator to explore underlying patterns of a large collection of GME records using data mining techniques. We define PHI as a vector of scores, each of which is a compliment probability defined based on the health-related risks associated with a particular disease category. Since the highest health risk is health-related death, we explore the health-related main Cause of Death (COD) information linked to the GME participants. Based on this definition, the higher the scores, the healthier the person. It is our belief that medical decision support systems are used to *support* clinical professionals rather than to *replace* them. So the primary goal of the proposed MyPHI is to draw their attentions to participants with high risks.

To the best of our knowledge, this work is the first of this kind in predicting

personal health scores by mining large medical examination data. PHI provides an important benchmark for understanding health status of the elderly people. Particularly, the following parties can be benefited by PHI:

- **Governments:** Public health policies are often made and revised based on scientific evidence from statistical analysis and research outputs [5]. For example, community health index can help the understanding of regional health status [6]. Public health authorities can use PHI to gauge their decisions on population health policies by utilizing the aggregated PHI of individuals. Particularly, the impact of a policy on regional health can be tracked by studying the progression or fluctuations of PHIs in a given time period. In addition, the population health in different regions can be compared and contrasted using the accumulated PHIs.

- **Organizations** such as hospitals, insurance companies, and nursing homes can better understand health status of the elderly through PHI. A systematic review in 2012 [7] suggested that clinical decision support systems (CDSSs) generally improved healthcare process for preventive and other types of services, although the strength of evidence is application dependent. Examples of recently developed data mining-based CDSSs include, but not limited to, coronary surgery management [8], drug prescription [9], cancer survivability [10] and nosocomial infections [11] predictions, and clinical monitoring in Intensive Care Unit (ICU) [12].

  At hospitals, our proposed PHI can assist physicians with additional information on patient conditions based on machine-extracted patterns across a range of disease specialities, which would not be otherwise obtainable via manual processing. Insurance companies may use the predicted PHI to assess personal health risks for consolidating insurance plans. The value of such predictions was seen in the 3-million prize for hospitalization prediction offered by the Heritage Provider Network in 2012 [13]. With PHI, nursing home carers can easily track elderly people's health status, and adjust custodial and medical care plans accordingly.

- **Individuals** can be benefited for having a peace of mind about their health status. People take regular medical examinations mostly not for discovering diseases but for obtaining a peace of mind regarding their health status. As a general practice, participant will have to meet a general practitioner (GP) after the examination for receiving and understanding the results [14], based on which the GP might suggest lifestyle change, further screening, or appointment with specialists. Once PHI is adopted, it can come with the report and provide additional information by revealing the person's health risks in various disease categories. It can be useful for motivating a participant with low PHI to follow doctor suggestions. In addition, the yearly PHIs can be used to track personal health status.

Indices or scores calculated based on patient conditions are commonly used in clinical practice, for example the severity of illness scoring systems for Intensive Care Unit (ICU) [15] and survival prediction tools for palliative care [16]. These tools often serve as a starting point for clinical diagnosis or prognosis [15, 17]. However, existing studies are generally based on statistical analysis on a small set of factors manually selected by medical experts, but in an era of information explosion, it is no longer possible to process all the information available and select factors manually.

In the last decade, increasing number of data mining applications has been developed to support healthcare decision making [18, 19]. Of one particular focus in recent years is the clinical risk classification [20, 21, 22, 12, 23]. These studies generally treat class labels with 100% certainty. However, label uncertainty is commonly found in clinical judgments due to expert subjectivity and inadequate information [24]. Often it is handled as noise, and the task is to detect and correct mislabeling [25, 26, 27]. However, in the case of multiple, non-exclusive medical conditions [28], such as comorbidity, it makes more sense to treat labels with degrees of certainty rather than forcing them to belong to one "true" class.

Recently in the field of Computer Vision, Yi *et al.* introduced a soft-label

learning model for complex event detection on Web videos [29]. For a given small number of target event instances, the model leverages related instances whose "relativeness" is uncertain and to be learned. Inspired by this work, we formulate the PHI prediction problem as a soft-label optimization problem. In the process of soft-label learning, we distinguish three types instances, namely participants with a target COD label, participants with a non-target COD label, and participants without a COD label. Traditionally in a binary prediction problem on a target event, instances of the first type are regarded as positive and those of the other types are treated as negative. However, in our case, participant records with different COD labels might share similar traits due to comorbidity, so some non-target instances could be "related" to the target ones. Our proposed method can capture the differences amongst these three types of instances.

This article substantially improves our previous work presented in a conference [30]. Firstly, we extend the concept of PHI from a single overall health score based on all-cause mortality to a vector of scores, each reflecting personal health risk in a disease category. Secondly, rather than treating labels as 100% certain, as in our previous work, we take a soft-label learning approach to handle label uncertainty. Experimentally we demonstrate the effectiveness of MyPHI based on a large geriatric medical examination (GME) data set of 262,424 records from 102,258 participants.

The rest of the paper is organized as follows. Section 2 describes our data sets. Section 3 details our MyPHI method and highlights the optimization technique employed to construct the prediction model. In Section 4, we demonstrate the effectiveness of our method through extensive experiments. The visualization of PHI is demonstrated in Section 5. In Section 6 we review the existing health scoring systems and models that handle label uncertainty. Section 7 concludes our work and discusses the further research directions.

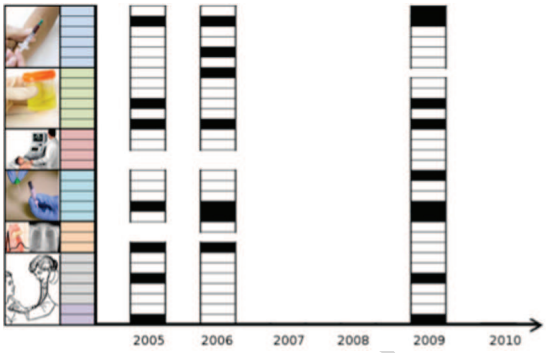


Figure 1: An example of a person's medical examination records over years. The vertical axis indicates the test items of various categories. Cells in black mark the items with abnormal results. The years without records, say 2007, 2008 and 2010, are the years that the person did not take the examination.

2. Data collection

Two data sets used in this study are a geriatric medical examination data set and a cause of death data set, linked together via the common attribute Person ID, revealing the associations between examination results and main causes of death.

2.1. Geriatric medical examination (GME) data set

GME is a de-identified data set with all private information, such as name, contact detail, and birth date, removed. Our GME data set has 230 attributes, containing 262,424 check-ups of 102,258 participants aged 65 or above, collected in a period of six years (2005 - 2010). Each de-identified GME 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 assessments, the Short Portable Mental Status Questionnaire (SPMSQ) cognitive function assessments, (de-identified) demographics as well as personal health-related habits, such as exercise, eating, drinking, and smoking habits. Key attributes in the above-mentioned categories are listed in Table 1. An example of a participant's GME records in years is included in Fig. 1. The person took three non-consecutive years of GMEs within a six-year period and

the abnormal results were marked black. Due to the voluntary nature of GME programs, the average number of records per participant is 2.56.

Three types of information were recorded for laboratory test and some physical examination items, namely the numerical results, status and descriptions. Numerical results are the machine output values of the test samples. Status shows if the values are normal based on the reference value range of the machine. Description fields give further text explanation on the abnormal results. As the ranges of reference values differ in different hospitals and such information is not given, numerical results are not comparable either between participants or for the same participant, because she might choose different hospital in different years. Therefore, only the status values are used in this study.

The data were collected in a standard annual medical 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 were removed before the data acquisition. The acquisition and processing of the data were approved by the Institutional Review Board (IRB) of the Taipei City Hospital.

2.2. Cause of death (COD) data set

The Cause of Death (COD) data set stores the cases that people who had been taking geriatric medical examinations (GME) for some years and then passed away. These records are linked to the GME data set via the common attribute Person ID. The COD data set records the main causes of death of all Taipei citizens, encoded with the WHO International Classification of Diseases, 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 data set. Attributes available from the linked information include a 3-4 digit ICD code for main cause of death, and time of the death (month and year). For example, the "malignant neoplasm of bronchus and lung" is encoded as 162 in ICD-9 and C349 in ICD-10. In addition, only health-related CODs are considered in this

Table 1: Selected GME 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	fobt
Examinations	Physical	weight, height, waist, systolic, diastolic, pulse
	External	neck, chest, heart, breast, abdomen, back, rectum, limbs, prosta
	Other	X-ray, ECG, cervical smear, abdominal ultrasound
Mental Health	BSRS	5 questions on 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.

research. The number of participants in the GME data set with COD codes is 7,569, accounting for 7.4% of the participants.

2.3. Data Characteristics

We have identified three characteristics from our GME data set, namely infrequency, incompleteness, and sparsity.

- **Infrequency:** As GME is offered on a yearly basis, the record sequences are infrequent, compared with time-series medical data, such as ECG and body movements collected from wearable sensor devices [12, 31]. Time-series data from wearable devices are often collected in the frequency of Hertz, so the problem is often on how to extract more compact data representation to save computational cost [32]. By contrast, the infrequency of GME data gives us relatively short sequences.

- **Incompleteness:** Due to the voluntary nature of GME, a person may only take a couple of GMEs in his lifetime. For example, as shown in Figure 1, the person took GME in three non-consecutive years. Therefore, the record sequences are incomplete along the timeline.

- **Sparsity:** Clinical judgements are often based on abnormal findings such as observed symptoms, signs, or lab test results. From the perspective of abnormalities, GME records are sparse, because the majority of the results would be normal.

These three characteristics of GME data differentiate our work from two strains of traditional classification problems. Our problem is different from the traditional point-based classification, since a person may have more than one records. However, our problem is also different from the traditional time-series classification problem, which is often on high frequency series (e.g., ECG) [32, 31], due to the infrequency and incompleteness of our data.

3. The Methodology

The proposed MyPHI prediction method that computes Personal Health Index (PHI) for elderly healthcare contains three key components, namely data

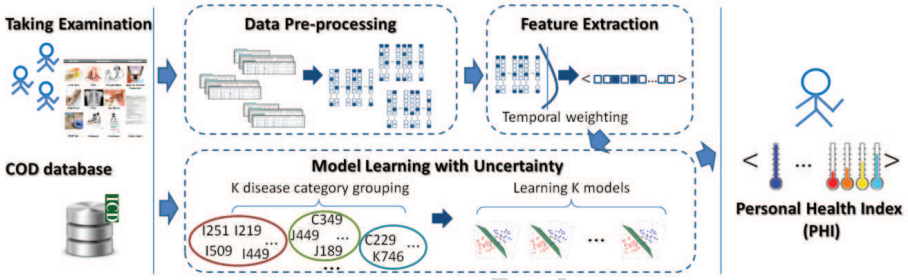


Figure 2: The process of Personal Health Index (PHI) prediction

pre-processing, feature extraction, and model learning with uncertainty, as shown in Fig. 2. The inputs of the method are the geriatric medical examination (GME) records of a population linked to the main cause of death (COD) database. The output is a vector of k predicted scores in the range of $[0, 1]$ interval, reflecting personal health risks in k disease categories. Note that the algorithm is designed for medical data sets that share the characteristics described in Section 2.3. Although in the following discussions we will use the GME data set as an example, the applicability of our proposed method is not limited to the data set.

3.1. Data pre-processing

Due to the noise observed in real-world medical data, the raw data need to be pre-processed. Firstly, we conducted data cleaning to remove extra or unrecognizable symbols, converted wide characteristics from Asian-based key-in systems into narrow characteristics, and corrected obvious misspellings. Secondly, for the free text fields, we only consider the "reasons for taking medicine" and "body obstacle type" fields of less noise. These texts were first tokenized and tokens with top frequency counts are extracted as additional binary features. Finally, since GME records are essentially longitudinal, participants with only one record were excluded. This leaves us a subset of 221,074 records from 60,881 participants.

In addition, we adopt an event-based view that treats an observed abnormal result as an occurrence of an event, following the practice of evidence-based medicine that only takes the observed symptoms and signs into consideration. For binary variable, abnormality is encoded as 1, and 0 otherwise. Real values are firstly discretized into bins. Ordinal and categorical variables are binarized into a vector of binary variables representing the unique values of the original variables. A variable takes the value 1 if the original variable takes the corresponding value; otherwise, it is 0.

After the pre-processing stage, participant longitudinal records can be represented as a sequence of time-stamped records, which can be formally defined as follows:

Definition 1 (Record sequence). A record sequence s_i of a person p_i is an ordered list of m records $\langle r_{i1}, ..., r_{ij}, ..., r_{im} \rangle$, where record r_{ij} is a tuple (t_{ij}, v_{ij}) of a d dimensional binary vector v_{ij} of values observed at time t_{ij}, where t_{ij} is the normalized time mapped onto an integer space such that $t_{ij} \in \mathbb{N}$ and $t_{ij} < t_{ij+1}$.

3.2. Feature extraction

One of the key challenges we face is how to represent a participant's records. To begin with, we take a feature-based approach [32] that converts sequences into a point-based representation, i.e., by transforming a sequence into a vector of features. The decision is based on the following reasoning. A participant can have multiple records as depicted in Fig. 1, so it is not naturally in a form of feature vector. In addition, these records cannot be simply flatten into a feature vector because record sequences of participants have varied lengths. In fact, record sequences are time-wise incomplete, as the average number of records per participant is 2.56 (Section 2.3). These considerations have led us to abandon the more intuitive sequence-based methods, so the problem becomes how to design a transformation mechanism that has a greater ability to capture different shapes of curves.

Our previous study on representation extraction strategies [30] suggests that time smoothing kernels that assign time weight to values at time t outperform methods without considering the longitudinal progression. Based on the previous results, we design a chi-squared kernel in current work to model the changes of importance over time. It is chosen over the commonly used Gaussian distribution because it has a greater ability in capturing different shapes of curves. The probability density function of the chi-squared distribution is defined as:

$$f(t,\theta) = \begin{cases} \frac{t^{\theta/2-1}e^{-t/2}}{2^{\theta/2}\Gamma(\theta/2)} & t \geq 0 \\ 0 & otherwise \end{cases} \tag{1}$$

where θ is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function.

The chi-squared kernel is defined as a function of truncated chi-squared distribution:

$$K_\theta(t) = \begin{cases} \frac{f(t,\theta)}{\Phi(T)-\Phi(1)} & t \in [1,T] \\ 0 & otherwise \end{cases} \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $f(\cdot)$.

The record sequence s_i of a person p_i is then transformed into a vector x_i using Eq.(2):

$$x_i = \sum_{j=1}^{m} K_\theta(T - t_{ij} + 1) \cdot v_{ij} \tag{3}$$

where the $(T - t_{ij} + 1)$ term reverses the time ordering, resulting in giving higher weights to latest records; v_{ij} is the j^{th} record of s_i obtained at time t_{ij}. Algorithm 1 shows the full procedure of feature extraction.

3.3. Model learning

To build an effective PHI prediction model for the elderly, we train k models, each for a target disease category. As discussed earlier in Section 1, there can be a degree of uncertainty that a COD is assigned to a person. This uncertainty can be observed in two types of instances, namely the target disease instances (known as the positive examples), and the non-target disease instances (usually treated as the negative examples). Intuitively, different people can belong to a target disease in different degrees. In addition, non-target cases and alive cases,

Algorithm 1 Feature Extraction

Input: S: a list of record sequences, θ: a scale parameter for the temporal weighting kernel, T: the time window of interest.

Output: X: extracted feature vector of S.

$X := []$

while $i < size(S)$ **do**

$((t_{ij}, v_{ij}))_{j \in 1...m} := S(i)$

$X(i) := \sum_{j=1}^{m} K_\theta(T - t_{ij} + 1) \cdot v_{ij}$

end while

return X

though all considered negative examples in the traditional sense, are "negative" in different ways: non-target cases may be closer to the target cases than to the alive cases.

3.3.1. Optimization problem

Due to the uncertainty of labels, we formulate the problem as an optimization problem [29] that finds a soft label for every instance. More specifically, given training instances $X = \{x_1, ..., x_n\}$ converted from n record sequences using Eq. (3), the soft label Y_i of x_i is $1 + S_i$ for target instances, $1 - S_i$ for non-target instances, and 0 for alive instances, where S_i that expresses the degrees of certainty is a variable to be learned. This design can be expressed by defining $Y = Y^a + A \odot S$, where $A \odot S$ is the entrywise product of vectors $A \in \mathbb{R}^n$ and $S \in \mathbb{R}^n,\ S \geq 0$. $Y_i^a = 1$ for the target and non-target instances and $Y_i^a = 0$ otherwise; $A_i = 1$ for the target instances, $A_i = -1$ for the non-target instances, and $A_i = 0$ otherwise.

So the optimization problem can be defined as a regularized least squared minimization problem with additional constraints on Y and S:

$$\begin{aligned} &\min_{P,S,Y} ||X^{\mathrm{T}} P - Y||_F^2 + \Omega(P) \\ &s.t.\ Y = Y^a + A \odot S,\ S \geq 0 \end{aligned} \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm such that $||B||_F = (\sum_{i=1}^{m}\sum_{j=1}^{n}|b_{ij}|^2)^{1/2}$ and $\Omega(\cdot)$ is a regularization term on P to prevent over-fitting.

We further postulate that noises of the labels are mostly from the non-target disease cases, as these cases could exhibit similar or completely different trait to the target cases. So a weight vector W is introduced, which is learned only using the target disease instances and alive instances. By applying trace norm $||\cdot||_*$ on $E = [W, P]$, P can be further constrained by W, as trace norm can uncover the shared knowledge of W and P [33]. The optimization problem is modified as:

$$\begin{aligned} \min_{W,P,S,Y} & ||\tilde{X}^{\mathrm{T}}W - \tilde{Y}||_F^2 + ||X^{\mathrm{T}}P - Y||_F^2 \\ & +\alpha(||W||_F^2 + ||P||_F^2) + \beta||E||_* \\ s.t.\ & Y = Y^a + A \odot S,\ E = [W, P],\ S \geq 0 \end{aligned} \tag{5}$$

where $\tilde{X}$ is the input data for the target and alive cases only, and $\tilde{Y}$ is the corresponding labels: $\tilde{Y}_i = 1$ for the target cases and $\tilde{Y}_i = 0$ for the alive cases. W and P are regularized using the Frobenius norm to prevent over-fitting, and together (i.e., E) they are further regularized using the trace-norm. α and β are the coefficients for the regularization terms.

3.3.2. Optimization Procedure

Now we describe the procedure of solving the optimization problem formulated in Eq. (5). Let $D = \frac{1}{2}(EE^{\mathrm{T}})^{-\frac{1}{2}}$. Eq. (5) can be converted to:

$$\begin{aligned} \min_{W,P,S,Y} & ||\tilde{X}^{\mathrm{T}}W - \tilde{Y}||_F^2 + ||X^{\mathrm{T}}P - Y||_F^2 \\ & +\alpha(||W||_F^2 + ||P||_F^2) + \beta Tr(E^{\mathrm{T}}DE) \\ s.t.\ & Y = Y^a + A \odot S,\ E = [W, P],\ S \geq 0 \end{aligned} \tag{6}$$

Eq. (6) can be solved through iteratively updating P, W, and S until convergence, by setting their partial derivatives to zero one at a time and solving it accordingly. The convergence is proved in [29]. By setting the derivative of Eq. (6) w.r.t. P to 0, we have:

$$P = (XX^{\mathrm{T}} + \alpha I + \beta D)^{-1}XY \tag{7}$$

Again, by fixing P and setting the derivative of Eq. (6) w.r.t. W to 0, we get:

$$W = (\tilde{X}\tilde{X}^{\mathrm{T}} + \alpha I + \beta D)^{-1}\tilde{X}\tilde{Y} \tag{8}$$

Optimizing S is to solve the following problem:

$$\min_{S \geq 0} ||X^{\mathrm{T}}P - (Y^a + A \odot S)||_F^2 \tag{9}$$

Let $M = X^{\mathrm{T}}P - Y^a$. The problem becomes:

$$\min_{S \geq 0} ||M - A \odot S)||_F^2 \tag{10}$$

Finally, the optimal solution to Eq. (10) is obtained by:

$$S_{ij} = \max(M_{ij}/A_{ij}, 0) \tag{11}$$

As shown in Algorithm 2, the optimal solution to Eq. (6) is obtained by iteratively updating P, W, S with Eq. (8) - Eq. (10) until convergence.

Given an example x_t, the predicted score for a target disease is $P^{\mathrm{T}}x_t$. Let all the k models (i.e., each is a P) be stored in $\Lambda \in \mathbb{R}^{d \times k}$. The prediction function $f : \mathbb{R}^{d \times 1} \rightarrow [-1, 1]^{k \times 1}$ can be defined as $f(x_t) = \Lambda^{\mathrm{T}}x_t$.

3.4. PHI calibration

To convert the predicted scores output by Algorithm 2 into probabilities and allow a person's PHI to be comparable to that of others, we further employ a step of PHI calibration. First, the scores are z-normalized within the model outputs, namely $z_k = \frac{f_k - \mu_k}{\sigma_k}$, where f_k is the k^{th} score of the output vector, and μ_k and σ_k are the mean and standard deviation of the k^{th} model outputs.

Since those with high risks are in the extreme end of the spectrum, we employ a generalized extreme value distribution [34] $G(t) = e^{-[1-t]}$, which has a steeper growth when $t > 0$. So the final PHI calibration function for the k^{th} model outputs is $G_k(z_k) = e^{-[1-z_k]}$. Finally, the k^{th} score of PHI is the compliment probability, i.e., $1 - G_k(z_k)$.

Algorithm 2 Learning with Uncertain Labels

Input: $X \in \mathbb{R}^{d\times n}$: extracted d dimensional feature vectors of n participants, $\tilde{X} \subseteq X$: a subset of X of size m, containing only the target and alive cases, $\tilde{Y} \in \mathbb{R}^{m\times 1}$: the corresponding labels of $\tilde{X}$, $Y^a \in \mathbb{R}^{n\times 1}$ and $A \in \mathbb{R}^{n\times 1}$: parameters for learning uncertain labels.

Output: Optimized W, P, S.

Set $t = 0$ and initialize W, P randomly;

repeat

Compute D_t as: $D_t = \frac{1}{2}(E_t E_t^{\mathrm{T}})^{-\frac{1}{2}}$;

Update P_t according to Eq. (7);

Update W_t according to Eq. (8);

Compute $M_t = X^{\mathrm{T}} P_t - Y^a$;

Compute S_t by $S_{tij} = max(M_{tij}/A_{ij}, 0)$ (Eq. (11));

$t = t + 1$

until *Convergence*

4. Emperiments and Evaluation

Extensive experiments were conducted to evaluate MyPHI using a real-world GME data set described in Section 2.

4.1. Disease category grouping

We selected top 10 disease categories that have the highest frequency counts in the GME data set based on the linked Cause of Death (COD) labels encoded in ICD9 and ICD10, as shown in Table 2. The "Other" category is defined to contain all the rest health-related ICD codes not in the top 10 disease categories.

4.2. Experiment setup

We compared MyPHI with two typical classification methods commonly used in medical applications as baselines, namely linear support vector machines (LinSVM) and logistic regression (LR). We used LIBSVM [35] for the implementation of LinSVM and LIBLINEAR [36] for the implementation of LR. In

Table 2: Numbers of positive cases in disease categories

Top k	Disease Category	Count
1	Lung	649
2	Heart	296
3	Cerebrovascular	153
4	Diabetes	112
5	Stomach	105
6	Colon	101
7	Liver	83
8	Pancreas	61
9	Septicaemia	60
10	Hypertension	42
11	Other	1,314

addition, we compared MyPHI with the class-weighted versions of the baseline methods, denoted as LinSVM-W and LR-W respectively, where the class weights were set according to the ratio of positive and negative class. We trained a model for each of the 11 disease categories, following the 35:35:30 stratified train/validate/test split ratio in all experiments. The negative (alive) cases were sub-sampled according to the positive vs. negative ratios 1:1, 1:10 and 1:100. For the "Other" disease category in the case of 1:100, since the portion of negative size exceeds the total number of negative cases, we report only the results of 10 disease categories. The parameters of all the algorithms were searched on the grid of $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ using the validation set. The parameter θ for chi squared kernel in feature extraction phase was experimentally set to 4.

4.3. Results

4.3.1. Clean vs. noisy cases

The algorithms were firstly evaluated under the ideal situation, where there are only positive instances (i.e., those whose main cause of death is the target disease) and negative instances (i.e., those who are alive). We call this setting

Table 3: The averaged AUC (%) of 11 disease categories of various positive vs. negative ratios. The proposed MyPHI significantly outperforms all the other algorithms in most cases.

	Clean Case					Noisy Case				
Ratio	MyPHI	LinSVM	LinSVM-W	LR	LR-W	MyPHI	LinSVM	LinSVM-W	LR	LR-W
1:1	**83.48**	82.11	82.11	82.79	82.79	**74.07**	71.23	71.23	70.47	70.47
1:10	85.65	77.99	**85.66**	78.72	85.19	**85.14**	75.03	81.56	75.65	80.32
1:100	**89.95**	68.57	85.91	70.42	85.99	**89.37**	69.1	83.83	70.45	84.04

the clean case, as there is no non-target instances to confuse the learning algorithms. The results are listed on the left of Table 3 using the Area Under the receiver operating characteristic Curve (AUC) measure under various positive vs. negative ratios settings. It is clear that MyPHI outperforms all other methods in most cases, and is comparable with LinSVM-W under 1:10 ratio. In addition, it can be observed that baseline methods without considering class weighting perform poorly at ratio 1:100. On the other hand, the class-weighted versions of baselines (i.e., LinSVM-W and LR-W) can better handle class imbalance, though not as good as MyPHI. In fact, MyPHI achieves its highest performance at 89.95% averaged AUC under 1:100 ratio.

On the right of Table 3, we compare the algorithms in a noisy case, where the non-target instances are introduced by sampling the same amount as the target instances. It can be seen that the performance is greatly compromised in the case of 1:1 ratio for all algorithms. This shows that non-target cases do confuse algorithms given limited learning instances. However, when the portion of negative instances increases, the performance bounds back. This may explain how larger training instances can help mitigate noise. In addition, the gap between MyPHI and other methods is enlarged in the noisy case. In fact, the performance of LinSVM-W and LR-W drops significantly in the noisy case. These results demonstrate the robustness of our method.

4.3.2. Individual disease categories

We further compared the results at the level of individual disease category. Fig. 3 shows the performance of the clean case in the upper graph and noisy case in the lower graph. It can be seen that although performance varies according

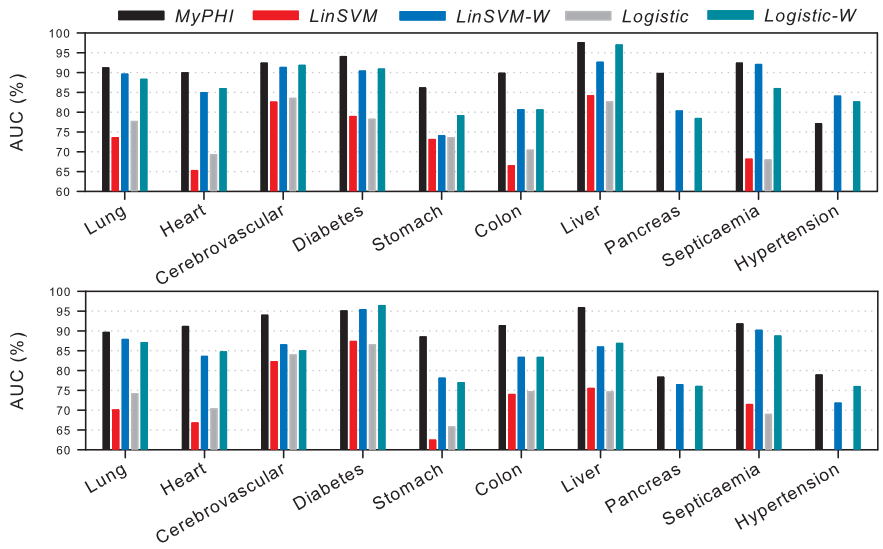
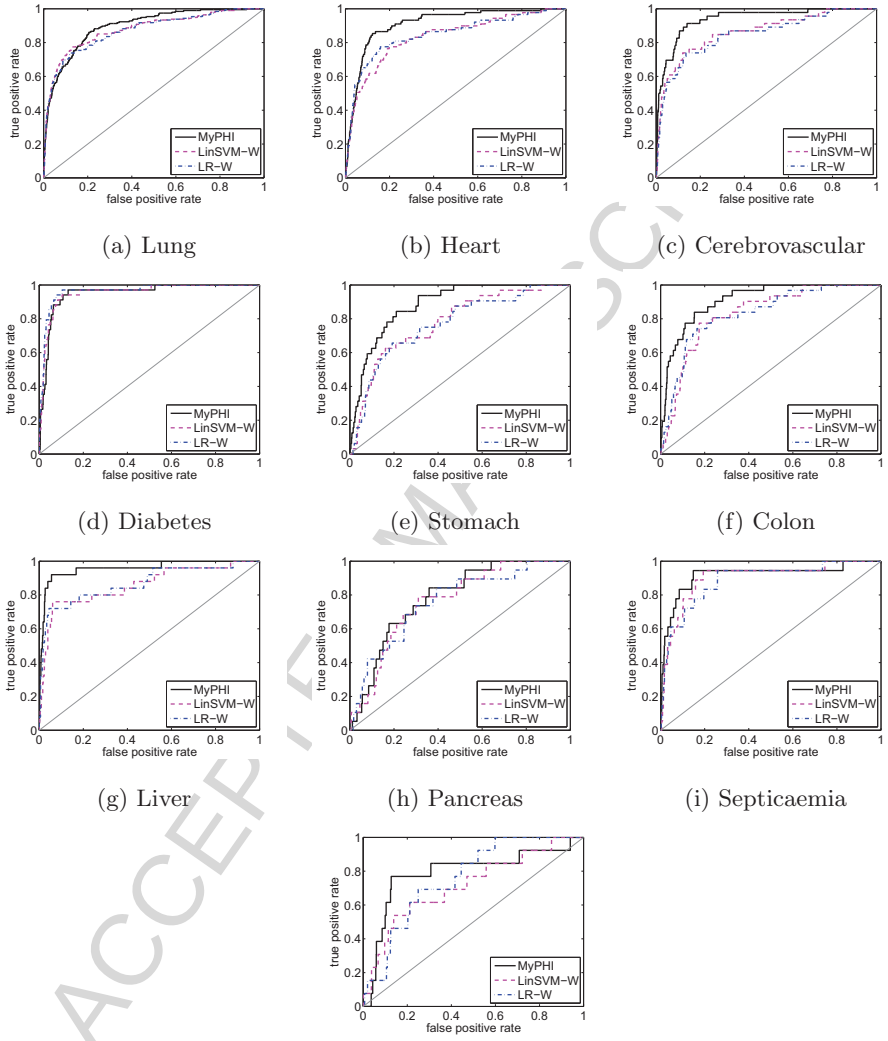


Figure 3: The AUC for individual disease categories under the 1:100 positive vs. negative ratio: the upper graph shows the performance in the clean case, where no instances of non-target disease categories are present; the lower graph displays the performance in the noisy case, where instances of non-target disease categories are introduced.

to the categories, MyPHI generally outperforms the other methods. One exception is Hypertension in the clean case; however, our method performs better in the noisy case. Another exception is Diabetes in the noisy case. For Heart, Cerebrovascular, Stomach, Colon, Liver, and Hypertension disease categories, MyPHI shows significantly better results than the other methods.

We looked into the Receiver Operating Characteristic (ROC) curves of the results in individual disease categories. Fig. 4 compares the performance of MyPHI, LinSVM-W and LR-W in the noisy case under 1:100 positive vs. negative ratio. The disease categories are ordered as before according to their sizes and the results for top 10 categories are displayed. MyPHI clearly dominates LinSVM-W and LR-W on the ROC graphs in the cases of Heart, Cerebrovascular, Stomach, Colon, Liver, and Septicaemia categories, while MyPHI is com-



(a) Lung (b) Heart (c) Cerebrovascular

(d) Diabetes (e) Stomach (f) Colon

(g) Liver (h) Pancreas (i) Septicaemia

(j) Hypertension

Figure 4: The ROC of top 10 disease categories under 1:100 positive vs. negative ratio and noisy settings.

parable to the two for Lung, Diabetes, and Pancreas categories.

4.3.3. Effects of trace norm

We also investigated the effects of introducing the trace norm constraint on W and P (Eq. 5), where target and alive cases are used to regulate the less

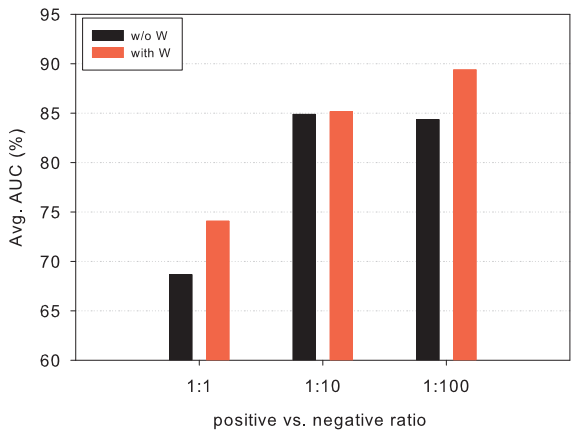


Figure 5: Comparing the effects of applying the trace norm in averaged AUC: "with W" denotes Eq. 5 where W is introduced and trace norm applied, while "w/o W" refers to Eq. 4 without the effect of trace norm.

certain labels from the non-target cases. As shown in Fig. 5, introducing W and trace norm improves the algorithm's performance. The effects are more significant in the cases of positive vs. negative ratio 1:1 and 1:100.

4.3.4. Effects of record sequence length

Studies have shown that the incompleteness of data can degenerate prediction performance [37]. As discussed earlier in Section 2.1 that the averaged record sequence length, i.e., number of records per person, for our data set is 2.56, we further conducted experiments to investigate the effects of record sequence length on the performance. Fig. 6 shows the averaged AUC under 1:100 ratio with the 95% confidence limits as the error bars. The Standard Error of the Mean (SEM) is used to calculate the standard error, i.e., $\mathrm{STD}/\sqrt{\mathrm{n_l}}$, where STD is the standard deviation and n_l is the number of cases with record sequence length l. The upper and lower confidence limits can be calculated as $\bar{x} \pm 1.96 \times \mathrm{SEM}$, where $\bar{x}$ is the mean AUC.

Fig. 6 shows that predictions with 5 records have the highest averaged AUC, followed by those with 3 and 4 records. Predictions with 2 records have the

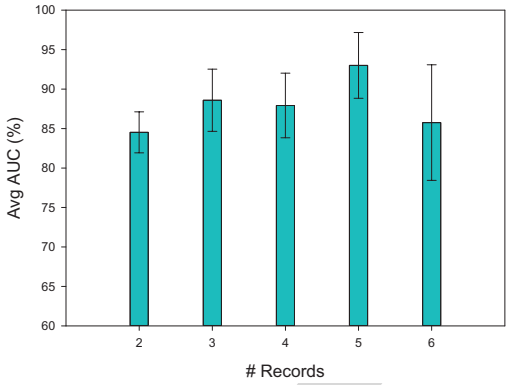


Figure 6: Comparing the effects of the number of records per person on the performance in averaged AUC. The error bar is calculated by Standard Error of the Mean (SEM).

lowest averaged AUC score of 84.53%, while the averaged AUC for predictions with 6 records has a larger SEM due to insufficient cases in some runs.

4.4. Discussions

Label uncertainty is often observed in real-world medical data. Our extensive experiments on a large GME data set with 262,424 check-ups of 102,258 participants have shown the robustness of our model under label uncertainty and class imbalance. Specifically, MyPHI achieves its best performance at above 89% averaged AUC under 1:100 ratio, about 5% higher than LinSVM-W and LR-W (Table 3), two de facto standard classifiers. The best AUC for a single disease category is achieved in the prediction under the Lung-related disease category at 96.95% AUC (Fig. 3). The ROC curves (Fig. 4) further confirm the above analysis at a fine-grand level. These results suggest that the proposed soft-label learning model is able to better handle label-uncertain data commonly found in medical applications. The parameter S for soft-labeling allows the algorithm to learn the degree of certainty for every data instance based on the data structure.

In addition, our experiments show that introducing the terms for W and applying the trace norm on $E = [W, P]$ help regularize P, with which the performance is improved (Fig. 5). This suggests that in dealing with uncertainty,

expert knowledge on what kind of labels may be more certain than others is important for designing an effective model.

Finally, we compare the effects of the lengths (2 to 6) of the record sequences on the performance in Fig. 6. Predictions based on 5-record sequences have the best performance, while those with 2 records have the lowest. The corresponding confidence limits based on the computed mean and standard error can be taken into account along with the predicted PHI given the record sequence length.

Note that the current model is built based on the GME data set, so it is limited to the associated population, namely the elderly residents of Taipei City. However, the proposed methodology is not limited to the current data set, but can be applied to other medical data sets with similar characteristics as stated in Section 2.3. We are also aware of the possible misinterpretation of the predicted results by non-professionals, which could happen to any screening results [38]. So the understanding and interpretation of the predicted PHI scores need to be assisted by a general practitioner.

5. PHI visualization

We provide a visualization interface that allows health state analysis at the personal level, as well as the population level.

5.1. Personal health analysis

Health analysis at a personal level provides insights into individual's health conditions [39, 40] and lays the foundation of effective health monitoring [31, 12]. With MyPHI, a person's PHI scores in disease categories can be computed, and the results can be displayed as a fingerprint chart on the left of Fig. 7, where an annulus represents the PHI scores of a year and a colored cell reflects the degree of severity for a disease category in that year. The white annuluses denote the years with no examinations taken. Yearly PHI scores of a particular disease category can be compared in the top-right chart for trend analysis. This visualization of PHI can make it easier for clinical professionals to grasp a person's health status.

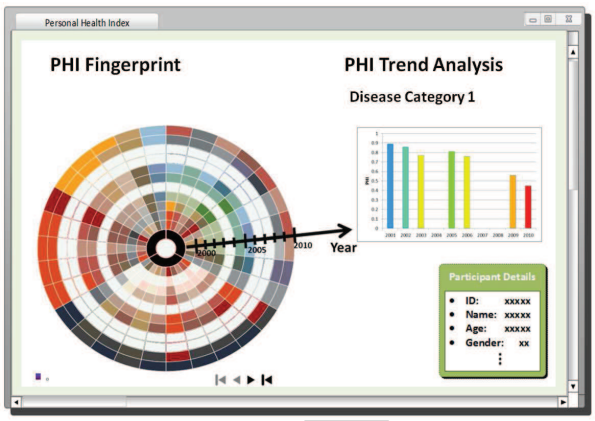


Figure 7: A dashboard of personal health analysis. The fingerprint on the left gives the breakdown PHI scores in disease categories, where an annulus of the concentric circle represents the PHI scores of a year and a sector in an annulus denotes the score of a particular disease category of that year. Color indicates severity and the white annuluses denote no attendance. In addition, yearly PHI scores of a disease category are summarized in the bar chart on the right.

5.2. Population health analysis

Health analysis at a population level can reveal regional health conditions and assist local government's fine-tune health policy making [41, 42, 4, 43, 44]. Fig. 8 shows a dashboard of regional health based on the averaged PHI scores of a disease category for the 12 districts of Taipei City, indexed by their zip codes. The colors on the district map reflect the degree of regional health using a color spectrum from red to green denoting PHI in the [0.5,1.0] range. Note that since the current model is trained based on the senior population of Taipei City, the use of the computed PHIs should be limited to this context.

6. Related work

6.1. Existing health scoring systems

Health indices provide numerical expressions of health status [45]. Many scoring systems were introduced to assist clinical decision-making, for example, the APACHE, SAPS, and MPM for Intensive Care Unit patients [15] and

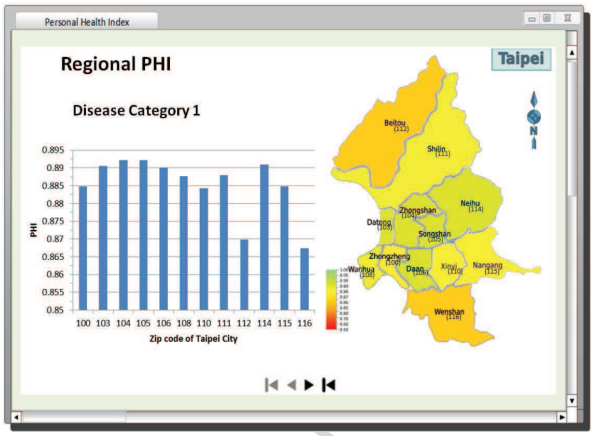


Figure 8: A dashboard of Taipei City's regional health based on the averaged PHI scores of a disease category for the 12 districts indexed by their zip codes. The colors on the district map reflect the degree of healthness based on the PHI scores.

the survival prediction tools for palliative care [16]. A systematic review conducted in 2012 [7] showed that clinical decision support systems (CDSSs) improved healthcare process for preventive and other types of services. Generally, these methods are defined based on factors selected with expert knowledge and validated via population-based studies [46]. However, as discussed earlier, it becomes problematic when the dimensionality increases and the longitudinal aspect is involved. Yi *et al.* [47] developed a bio-mark based system to grade personal health status. Again the model relies on the experts to define factors and the associated weights. Recently, Rothman *et al.* filed a patent for a monitoring system that computes patient health scores based on Electronic Health Records (EHRs) [17]; however, the underlying scoring mechanism is unknown.

In our early study [30], we introduced a classification-based framework that predicts Personal Health Index (PHI) as an overall health score to provide feedback to individuals based on the evaluation of risks revealed in their medical examination records. A binary Support Vector Machine classification model was built based on the same GME data set as described here in Section 2. However, the linked cause of death (COD) information was used as binary labels,

positive for the deceased cases and negative for the alive cases. The output is a predicted overall health score computed as the complement probability that a person belongs to the high risk class. However, the label uncertainty issue was not present in this binary classification problem setting, because live/dead are generally regarded as labels with high certainty [28].

6.2. Learning with label uncertainty

Label uncertainty is often treated as label noise in the literature [48]. In the medical domain, key sources of label noise are expert subjectivity and inadequate information in clinical judgments such as diagnosis [24]. Such noise is regarded as mislabeling to be detected and corrected [25, 26, 27]. For example, Garca-Zattera *et al.* employed binary Markov models to estimate misclassification parameters for dental research [27]. Rantalainen *et al.* introduced a Bayesian approach to detect control subjects who might be actually the undiagnosed cases in a case-control study [26]. The underlying assumption of these approaches is that ground-truth labels exist and are certain, so the task is to detect and correct mislabeling caused by human inadequacy.

However, uncertainty may come from multiple, non-exclusive medical conditions [28]. For example, health-related death can be caused by the complications of several co-existing diseases, and the identified main cause of death (COD) may only explain the death to a certain degree. In such cases, rather than identifying "mislabeled" cases, it makes more sense to allow labels to have degrees of certainty.

This problem can be resolved by adopting the soft-label learning approach introduced by Yi *et al.* in the field of Computer Vision for complex event detection [29] on Web videos. In the context of event detection on videos, there are often only a few positive instances available for training. On the other hand, there are related instances whose "relativeness" to the target event are uncertain. They formulated the learning problem as a soft-label optimization problem and demonstrated that exploiting related examples improved performance.

7. Conclusion

Computing comprehensive health scores for citizens was not considered practical before the *big data* era. Because of the availability of a large volume of data collected from multiple sources with all kinds of methods over years, effectively evaluating health status of a person from-cradle-to-grave is becoming possible. One of such data sources is from the annual geriatric medical examination (GME) which is now an integral part of elderly healthcare for many developed countries. Predicting personal health status based on medical examinations reveals a promising and important trend in healthcare research.

In this article we described MyPHI, a data mining-based method that predicts Personal Health Index (PHI) based on GME records. The extensive experiments on a real-world GME data set of 262,424 records from 102,258 participants demonstrated that our model outperformed the commonly-used classifiers, such as linear SVM, logistic regression and their class-weighted versions. In particular, MyPHI is shown to be robust under label uncertainty and class imbalance, and achieved 89.95% averaged AUC (Area Under the receiver operating characteristic Curve) under a ratio of 1:100 positive vs. negative.

Overall, we believe we have provided a new direction of quantifying personal health through data mining techniques. We foresee three aspects to improve the PHI prediction in future.

- **Data semantics.** Text fields in GME records, such as doctor suggestions and result descriptions, may contain useful information that can be further incorporated into the evaluation to improve the effectiveness.

- **Data expansion.** Other information sources, such as online medical information, bio-marks collected from mobile sensor devices (e.g., accelerometer and gyroscope on a smart phone recording the daily activities of a person), and in-hospital data, can be linked to obtain more comprehensive results.

- **Dynamic system.** In current work, PHI is computed based on the

archived GME database, which can be extended to an online dynamical version.

Acknowledgment

We thank Dr. John Bennett at UQ Health Service for his insightful comments. This study is performed based on the de-identified data from the Taipei City Public Health Database at 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, Taipei City Government, or DoPHA. The research is partially founded by the Australian Research Council Discovery Project ID DP140100104.

References

[1] L. Mirel, K. Carper, Statistical Brief - Trends in Health Care Expenditures for the Elderly, Age 65 and Over, Tech. rep., Agency for Healtcare Research and Quality (2014).

[2] Health assessment for people aged 75 years and older, http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare_mbsitem_75andolder, accessed: 2015-06-08.

[3] Health checks for the over-65s, http://www.nhs.uk/Livewell/Screening/Pages/Checksover65s.aspx, accessed: 2015-06-08.

[4] Health Promotion Administration, Ministry of Health and Welfare, Taiwan, 2014 Annual Report of Health Promotion Administration Health, Available from http://www.hpa.gov.tw/English/ClassShow.aspx?No=201412160001 (2014).

[5] Understanding evidence-based public health policy, American Journal of Public Health 99 (9) (2009) 1576–1583. doi:10.2105/AJPH.2008.156224.

[6] Aggregate Health Status: a benchmark index for community health., Journal of Medical Systems 27 (2) (2003) 177–89.

[7] T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty, L. Wing, A. S. Kendrick, G. D. Sanders, D. Lobach, Effect of Clinical Decision-Support Systems, Annals of Internal Medicine 157 (1) (2012) 29–43.

[8] D. Delen, A. Oztekin, L. Tomak, An analytic approach to better understanding and management of coronary surgeries, Decision Support Systems 52 (3) (2012) 698–705. doi:10.1016/j.dss.2011.11.004.

[9] M. T. Akçura, Z. D. Ozdemir, Drug prescription behavior and decision support systems, Decision Support Systems 57 (1) (2014) 395–405. doi:10.1016/j.dss.2012.10.045.

[10] H. M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decision Support Systems 74 (2015) 150–161. doi:10.1016/j.dss.2015.04.003.

[11] M. Ben Ayed, H. Ltifi, C. Kolski, A. M. Alimi, A user-centered approach for the design and implementation of KDD-based DSS: A case study in the healthcare domain, Decision Support Systems 50 (1) (2010) 64–78. doi:10.1016/j.dss.2010.07.003.

[12] Y. Mao, W. Chen, Y. Chen, C. Lu, S. Louis, M. Kollef, T. C. Bailey, An integrated data mining approach to real-time clinical monitoring and deterioration warning, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Beijing, China, 2012, pp. 1140–1148.

[13] Heritage provider network health prize, https://www.heritagehealthprize.com/c/hhp, accessed: 2015-06-08.

[14] Process of elderly health examination (in Chinese), http://w3srv.ntuh.gov.tw/bh/%E5%8C%97%E8%AD%B7%E5%81%A5%E6%AA%A20402/item.htm, accessed: 2015-05-20.

[15] A. Fisher, D. Burke, Critical care scoring systems, in: Contemporary Coloproctology, Springer, 2012, pp. 513–528.

[16] P. A. Glare, C. T. Sinclair, Palliative medicine review: prognostication, Journal of Palliative Medicine 11 (1) (2008) 84–103.

[17] M. J. Rothman, S. I. Rothman, D. B. Rothman, Systems and methods for providing a health score for a patient, US Patent 8,403,847 (Mar. 2013).

[18] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, V. K. Tabar, Knowledge discovery in medicine: Current issue and future trend, Expert Systems with Applications 41 (9) (2014) 4434–4463.

[19] J.-Y. Yeh, T.-H. Wu, C.-W. Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decision Support Systems 50 (2) (2011) 439–448.

[20] F. Wang, P. Zhang, B. Qian, X. Wang, I. Davidson, Clinical risk prediction with multilinear sparse logistic regression, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, USA, 2014, pp. 145–154.

[21] T. Tran, D. Phung, W. Luo, S. Venkatesh, Stabilized sparse ordinal regression for medical risk stratification, Knowledge and Information Systems (2014) 1–28.

[22] J. Wiens, E. Horvitz, J. V. Guttag, Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task, in: Neural Information Processing Systems, 2012, pp. 476–484.

[23] H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, M. Rosen-Zvi, 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, ACM, San Diego, CA, USA, 2011, pp. 395–403.

[24] M. Pechenizkiy, a. Tsymbal, S. Puuronen, O. Pechenizkiy, in: 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), IEEE, pp. 708–713.

[25] A. Malossini, E. Blanzieri, R. T. Ng, Detecting potential labeling errors in microarrays by data perturbation, Bioinformatics (Oxford, England) 22 (17) (2006) 2114–2121.

[26] M. Rantalainen, C. Holmes, Accounting for Control Mislabeling in CaseControl Biomarker Studies, Journal of Proteome Research (2011) 5562–5567.

[27] M. J. García-Zattera, T. Mutsvari, a. Jara, D. Declerck, E. Lesaffre, Correcting for misclassification for a monotone disease process with an application in dental research., Statistics in Medicine 29 (30) (2010) 3103–3117.

[28] L. Lin, P. J.-H. Hu, O. R. Liu Sheng, A decision support system for lower back pain diagnosis: Uncertainty management and clinical evaluations, Decision Support Systems 42 (2) (2006) 1152–1169.

[29] Y. Yang, Z. Ma, Z. Xu, S. Yan, A. G. Hauptmann, How related exemplars help complex event detection in web videos?, in: IEEE International Conference on Computer Vision (ICCV'13), IEEE, 2013, pp. 2104–2111.

[30] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, M. Sharaf, Mining personal health index from annual geriatric medical examinations, in: IEEE 14th International Conference on Data Mining (ICDM'14), IEEE, 2014.

[31] A. Pantelopoulos, N. G. Bourbakis, A survey on wearable sensor-based systems for health monitoring and prognosis, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 40 (1) (2010) 1–12.

[32] Z. Xing, J. Pei, E. Keogh, A brief survey on sequence classification, SIGKDD Explor. Newsl. 12 (1) (2010) 40–48.

[33] Y. Yang, Z. Ma, A. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Transactions on Multimedia 15 (3) (2013) 661–669.

[34] E. J. Gumbel, Statistics of extremes, Courier Dover Publications, 2012.

[35] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[36] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, The Journal of Machine Learning Research 9 (2008) 1871–1874.

[37] W. R. Hersh, M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. J. Cimino, J. H. Saltz, Caveats for the use of operational electronic health record data in comparative effectiveness research, Medical Care 51 (8 Suppl 3) (2013) S30–7. doi:10.1097/MLR.0b013e31829b1dbd.

[38] D. a. Grimes, K. F. Schulz, Uses and abuses of screening tests, Lancet 359 (9309) (2002) 881–884.

[39] F. S. Collins, H. Varmus, A New Initiative on Precision Medicine, The New England Journal of Medicine 372 (9) (2015) 793–795. doi:10.1056/NEJMp1415160.

[40] A. Klemes, R. E. Seligmann, L. Allen, M. a. Kubica, K. Warth, B. Kaminetsky, Personalized preventive care leads to significant reductions in hospital utilization, American Journal of Managed Care 18 (12) (2012) 4–5.

[41] U.S. Department of Health and Human Services, CDC's Vision for Public Health Surveillance in the 21st Century, Available from http://www.cdc.gov/mmwr/pdf/other/su6103.pdf (2012).

[42] Health information privacy - public health, http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/publichealth/index.html, accessed: 2015-05-20.

[43] D. J. Friedman, R. G. Parrish, D. a. Ross, Electronic health records and US public health: Current realities and future promise, American Journal of Public Health 103 (9) (2013) 1560–1567. doi:10.2105/AJPH.2013.301220.

[44] R. Kukafka, J. S. Ancker, C. Chan, J. Chelico, S. Khan, S. Mortoti, K. Natarajan, K. Presley, K. Stephens, Redesigning electronic health record systems to support public health, Journal of Biomedical Informatics 40 (4) (2007) 398–409. doi:10.1016/j.jbi.2007.07.001.

[45] R. Kaplan, J. Bush, C. Berry, Health status: types of validity and the index of well-being, Health Services Research 11 (4) (1976) 478–507.

[46] M. Woodward, Epidemiology: study design and data analysis, CRC Press, 2013.

[47] S.-I. Yi, B.-R. So, C.-S. Lee, S.-J. Lee, S.-K. Park, B.-K. Park, I.-W. Chung, Classification of Health Grade Using Bio-Check Unit and Health Index, Journal of Biomechanical Science and Engineering 6 (3) (2011) 148–159.

[48] B. Frenay, M. Verleysen, Classification in the presence of label noise: A survey, IEEE Transactions on Neural Networks and Learning Systems 25 (5) (2014) 845–869.

Biographical Notes

Ling Chen is currently a PhD Candidate in Computer Science at The University of Queensland (School of Information Technology and Electrical Engineering), Australia. Her research focuses on data mining and its applications in health care.

Xue Li is currently an Associate Professor at The University of Queensland (School of Information Technology and Electrical Engineering), Australia. He is an Adjunct Professor of University of Electronic Science and Technology, China and a Guest Professor of Chongqing University. His research activities mainly focus on Data Mining, Multimedia Data Security, Database Systems, and Intelligent Web Information Systems.

Yi Yang received the Ph.D degree in Computer Science from Zhejiang University, Hangzhou, China, in 2010. He is now a DECRA fellow with the University of Queensland, Brisbane, Australia. Prior to that, he was a Postdoctoral research fellow at the school of computer science, Carnegie Mellon University, Pittsburgh, PA. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g. multimedia indexing and retrieval, surveillance video analysis, video semantics understanding, etc.

Hanna Kurniawati is a Lecturer at The University of Queensland (School of Information Technology and Electrical Engineering), Australia. She received a Ph.D. in Computer Science from National University of Singapore. Her current research interest includes robotics, motion planning, planning under uncertainty, computational geometry applications, machine learning, and randomized algorithm.

Quan Z. Sheng is an Associate Professor and Deputy Head of the School of Computer Science at the University of Adelaide. He holds a PhD degree in computer science from the University of New South Wales, Australia and did his post-doc as a research scientist at CSIRO ICT Centre. Dr. Sheng 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). His research interest includes Web of Things, Big Data Analytics, Web Science, Service-oriented Computing, Pervasive Computing, Sensor Networks.

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 & Management from the Johns Hopkins Bloomberg School of Public Health. Her research interest includes Health Policy and Management, Health Services Research, and Health Disparities.

Highlights

- A new approach to predicting health scores based on medical examinations using data mining algorithms
- Quantitative analysis of elderly health status to support health management for the governments, organizations, as well as individuals
- Experiments performed based on a large and comprehensive geriatric medical examination data set of 102,258 participants