Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Identifying informative risk factors and predicting bone disease progression via deep belief networks

Hui Li [a,*], Xiaoyi Li [a], Murali Ramanathan [b], Aidong Zhang [a]

[a] Department of Computer Science and Engineering, State University of New York at Buffalo, USA
[b] Department of Pharmaceutical Sciences, State University of New York at Buffalo, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Informative risk factors
Osteoporosis prediction
Deep belief networks (DBNs)
Restricted Boltzmann machine (RBM)
Bone fracture

ABSTRACT

Osteoporosis is a common disease which frequently causes death, permanent disability, and loss of quality of life in the geriatric population. Identifying risk factors for the disease progression and capturing the disease characteristics have received increasing attentions in the health informatics research. In data mining area, risk factors are features of the data and diagnostic results can be regarded as the labels to train a model for a regression or classification task. We develop a general framework based on the heterogeneous electronic health records (EHRs) for the risk factor (RF) analysis that can be used for informative RF selection and the prediction of osteoporosis. The RF selection is a task designed for ranking and explaining the semantics of informative RFs for preventing the disease and improving the understanding of the disease. Predicting the risk of osteoporosis in a prospective and population-based study is a task for monitoring the bone disease progression. We apply a variety of well-trained deep belief network (DBN) models which inherit the following good properties: (1) pinpointing the underlying causes of the disease in order to assess the risk of a patient in developing a target disease, and (2) discriminating between patients suffering from the disease and without the disease for the purpose of selecting RFs of the disease. A variety of DBN models can capture characteristics for different patient groups via a training procedure with the use of different samples. The case study shows that the proposed method can be efficiently used to select the informative RFs. Most of the selected RFs are validated by the medical literature and some new RFs will attract interests across the medical research. Moreover, the experimental analysis on a real bone disease data set shows that the proposed framework can successfully predict the progression of osteoporosis. The stable and promising performance on the evaluation metrics confirms the effectiveness of our model.

1. Introduction

Osteoporosis is the most common type of bone diseases associated with aging and may be clinically silent but can cause significant mortality and morbidity after onset. It becomes an important public health issue because of costs associated with treatment interventions (increasingly including rehabilitation and extended treatment facilities), the chronic, prolonged course, high mortality rates following hip fracture, and high incidence in women. It is estimated to affect 44 million Americans. About 40–50% of women and 13–22% of men are at risk of having an osteoporotic fracture in their lifetime [1–3]. The care costs were estimated to be $19 billion annually in 2005, with an anticipated increase to $25.3 billion annually in 2025 [4–6]. The ability to leverage a quantitative paradigm to alleviate and improve patient outcomes, both in terms of diagnosis and prevention, would confer significant benefits both to individuals and to society.

Although the diagnosis of osteoporosis is usually based on the assessment of bone mineral density (BMD) using dual energy X-ray absorptiometry (DXA), BMD incompletely reflects the variation in bone strength and thus can not be used to monitor the disease progression. World Health Organization (WHO) has claimed that information integration on risk factors (RFs) is helpful on predicting the risk of bone disease in men and women worldwide [7]. Osteoporosis is an associated disease with potential RFs from various aspects such as demographic attributes, patients' clinical records regarding disease diagnoses and treatments, family history, and lifestyle. Usually, numerous potential RFs need to be considered simultaneously since observed and hidden reasons behind all RFs are worth learning for the exploration of the disease progression. However, it is an extremely challenging task to capture

* Corresponding author.
E-mail address: hli24@buffalo.edu (H. Li).

http://dx.doi.org/10.1016/j.ymeth.2014.06.011
1046-2023/

the disease characteristics and clinical nuances for predicting the disease progression and detecting the informative RFs due to the complexity and diversity of the data. These difficulties are reflected in at least two ways. First, it is hard to find a good set of RFs so that the salient integrated features can be disentangled from heterogeneous information. Second, it is difficult to discriminate the different roles of seemingly independent features for healthy patients and for diseased patients.

A variety of RF analysis models aimed at tackling with these challenges usually fall into two categories: the expert knowledge based model or the handcrafted feature set based model. The expert knowledge based model mainly relies on a small number of well-known RFs which have been validated by an expert in this field [8,9]. One of the most popular expert knowledge based models is WHO fracture risk assessment tool (FRAX) [9] which was developed by WHO and has been widely used to give the 10-year probability of osteoporosis and bone fracture. But it may not be appropriate to directly adopt the results since FRAX sometimes overestimates or underestimates the fracture risk [10]. Moreover, the information collected by FRAX is limited so that some important features might be discarded and thus affect the predictive performance. Thus, the prediction results from such a tool need to be further interpreted with caution and properly re-evaluated. The handcrafted feature set based model tries to find the informative RFs by calculating their statistical significance and then measure the predictive power. The assessment method of the relationship between a disease and a handcrafted risk factor is based on the regression model [11,8,12,13] such as linear regression, logistic regression, Poisson regression, Cox regression and other learning methods such as Artificial Neural Network (ANN) [14], association rules [15] and decision tree [16]. Although these models are theoretically acceptable for analyzing the risk dependence of several variables, it pays little attention to the relationships among RFs and between RFs and the target disease. Some methods require a good setting of meta parameters and so parameter-tuning is an inevitable issue. Furthermore, they usually extract the statistically significant features from a commonly known RF list, which means there still may be a loss of useful information if the list is not comprehensive. Mining the causality relationship between RFs and a specific disease has attracted considerable research attention in recent years. In [17–19], limited RFs are used to construct a Bayesian network and the RFs are assumed conditionally independent of one another since learning the Bayesian networks becomes tough and even impossible as the number of RFs increases. Some hybrid data mining approaches might also be used to combine classical classification methods with feature selection techniques for the purpose of improving the performance or minimizing the computational expense for a large data set [20], but they are limited by the challenge of explaining selected features.

Although the existing risk factor analysis methods provide us with meaningful interpretations of how to use selected RFs for disease prediction, there are still four major open problems: (1) the risk factors are usually constrained to a fixed number and are thus too simple to capture the comprehensiveness behind diseases, (2) selection of risk factors are specified by expert knowledge based on previous studies and thus might be inappropriate for new population samples, (3) methods are not robust in the presence of missing and noisy data, which is a common and challenging issue for medical data, and (4) the existing methods cannot be transferred to multiple tasks making it nearly impossible for generalization. To solve these problems, we develop an effective disease risk management model with a deep architecture. Recently, many efforts have been devoted to develop learning algorithms for the deep graphical model with impressive results obtained in various areas such as computer vision and natural language processing [21]. The intuition and extensive learning power of these models are suitable for our task. The main innovation of our proposed model is to find a representation of risk factors so that the salient integrated features can be disentangled from ill-organized data. Also, we try to discriminate the RF representation for both diseased and non-diseased individuals. Under this scenario, our model will be ultimately used to select the informative RFs and predict bone disease progression.

2. A novel risk factor analysis framework for bone health

We used a clinical data from study of osteoporotic fractures (SOF) [22]. The risk factors, as the input of our framework, are extracted from the baseline year for all participants. The outputs of the framework are the bone disease prediction results and a pool of informative risk factors. To evaluate the prediction results, we extracted the diagnostic results from longitudinal follow-up data after 10 years. The informative risk factors are examined by expert opinions. We used SAS software to retrieve data and Matlab software for data analysis, algorithm implementation and performance visualizations.

Fig. 1 shows the roadmap of the entire framework including EHR data collection, risk factor construction, data partition, model learning, informative risk factor selection and osteoporosis prediction. The description of each component is given as follows.

EHR data. The electronic health record (EHR) is a longitudinal electronic record of patient health information including diverse information like demographics, medications, past medical history, laboratory data, and lifestyles. EHRs are valuable sources for exploratory analysis and statistics to assist clinical decision-making and further medical research. In this paper, we use a public EHR data from the study of osteoporotic fractures (SOF) [22] which is one of the largest and most comprehensive study for bone diseases which includes 9704 Caucasian women and additional 662 African-American women aged 65 years and older. It contains 20 years of prospective data collected over nine completed visits about osteoporosis, bone fractures, breast cancer, and so on. Potential risk factors (RFs) and confounders were classified into 20 categories such as demographics, family history, lifestyle, and medical history.

Risk factor construction. We define the risk factors for all patients based on their EHR data. A number of potential RFs are grouped and organized during the baseline period, as shown in Fig. 2. After preprocessing the baseline data, 672 variables covering entire 20 categories will be the input of our model. Note that there are missing values for each patient, which is a common problem for EHR data. During the examination period (from the second visit), participants attended a series of examinations of BMD measured by DXA. We process these BMD data based on the WHO standard which will be used to define the diagnosis date, as shown in Fig. 2 with a star symbol. Specifically, *T*-score of less than -1[1] indicates the diagnosis of osteopenia[2]/osteoporosis. In Fig. 2, risk factors are constructed from the baseline period and we try to predict osteoporosis onset after the examination period. In fact, the earlier we predict osteopenia/osteoporosis before the diagnosis date, the better we can help patients prevent from getting worse.

Data partition. We partition the entire data into two subsets, in which one is used for learning our model and the other subset is used for testing. During the training phase, we further split data into diseased individuals and non-diseased individuals for training two separate models. During the testing phase, we apply mixed data in our model for evaluation.

[1] *T*-score of -1 corresponds to BMD of 0.82, if the reference BMD is 0.942 and the reference standard deviation is 0.122.

[2] Osteopenia is a pre-condition of osteoporosis. *T*-score of less than -2.5 is the indicator of osteoporosis.