Informatics in Medicine Unlocked 10 (2018) 134–142

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Data mining-based subhealth analysis of Chinese software programmers in 2017

Bingdi Wang [a,b,c], Shengqi Yang [a,*], Zhangqin Huang [a,b,c], Da Li [a,b,c], Jian He [a,b,c], Xiaoyi Wang [a,b,c]

[a] *Beijing Advanced Innovation Center for Future Internet Technology, China*
[b] *Beijing Engineering Research Center for IoT Software and Systems, China*
[c] *Beijing University of Technology, Beijing, China*

ARTICLE INFO

Keywords:
Chinese programmers
Subhealth
Data mining
Logistic regression

ABSTRACT

This work analyzes the suboptimal health (subhealth) status of Chinese software programmers in 2017 and reveals its major causes. By using the Chinese subhealth evaluation scale (CSHES), programmers from China were invited to participate in the designed survey. On the basis of the analysis of the programmer's score in terms of the CSHES, the data were processed with a logistic regression model and analyzed to reveal the major causes of subhealth. The data analysis results show that the Chinese programmers' subhealth issue in 2017 is mainly induced by metabolic disorders, depression, pressure, satisfaction, and the quality of their sexual life. Through this survey based on a data-mining study, Chinese programmers can clearly understand their own health status through their own subhealth characterization, discover their own possible causes of subhealth status, develop a healthy lifestyle in accordance with the corresponding symptoms, and improve their health level.

1. Introduction

With the continuous development of the social and medical economy and an accelerated improvement of science and technology level in China, people's awareness of health status has also changed in the 21st century. An increasing number of people are now focusing on health-related issues. These issues are related to people's daily lives and have a direct impact on their quality of life. Suboptimal health (subhealth), as a recently introduced concept of medicine, has attracted the attention of the research world. In addition to health and disease, there is a non-healthy and nondiseased intermediate state, the subhealth state, also known as the "third state," as proposed by the former Soviet Union Professor Berkman [1]. It refers to the individual in a clear diagnosis of physical, psychological, and social adaptation with a low-quality health status and experience [2]. If people who experience the subhealth status cannot get timely help and treatment, this status easily leads to physical and psychological diseases such as psychological disorders, gastrointestinal diseases, cardiovascular and cerebrovascular diseases, and cancer. Furthermore, subhealth status for people who are under extensive stress can also lead to sudden death, the so-called "death from overwork" [3], which is considered one of the main killers of human health in the 21st century. A national survey from the World Health Organization (WHO) shows that only 5% of people have a healthy status, 20% of people have illness, and about 75% of people are in a subhealth state [4]. In 2006, the China Association of Chinese Medicine divided subhealth into three subtypes: physical subhealth, psychological subhealth, and social subhealth [5,6]. Physical subhealth, according to the performance characteristics of the main symptoms, can be further divided into fatigue-type subhealth and pain-type subhealth. The fatigue subhealth, also known as chronic fatigue syndrome, is a common condition [7]. Most studies on subhealth from China targets research on chronic fatigue syndrome [8–10].

The etiology of subhealth is a very complicated problem. Recent studies suggest that the occurrence of subhealth can be related to personal physiological, psychological conditions, occupational circumstances, social environment, living environment, working methods, living habits, and other factors [11–14]. Health status is revealed by the evaluation of a wide range of symptoms, involving social, psychological, and physical manifestations, and other aspects, belong to the typical "massive data" characteristics of informatics with a variety of information complexity. As a result, it is difficult to conduct a general analysis [15]. At present, the application of data mining technology in the study

[*] Corresponding author. Beijing Advanced Innovation Center for Future Internet Technology, China.
E-mail address: syang@bjut.edu.cn (S. Yang).

https://doi.org/10.1016/j.imu.2018.01.001
Received 6 November 2017; Received in revised form 5 January 2018; Accepted 20 January 2018
Available online 31 January 2018

of subhealth status is mainly to obtain a large number of macroscopic features through the clinical epidemiological survey [16–18]. On the basis of this, traditional statistical methods are used to study the demographic characteristics of population with subhealth status, subhealth population classification, syndrome and symptom characteristics, subhealth status of individuals with different syndromes, influencing factors and risk factors, and subhealth quantitative assessment. Subhealth status-related data are quite different from other clinical data [19,20]. The subhealth status data has a large number of subjective symptoms, such as the direct measurement of some hidden health variables, and the use of the questionnaire to obtain a true and reliable reflection of subhealth traditional Chinese medicine syndrome characteristics. Furthermore, this kind of data evolves and changes through time [21]. It is therefore difficult to identify subhealth status-related indicators and organic syndrome types [22,23]. The use of traditional statistical methods is not enough to summarize the characteristics of the subhealth state.

In the present study, Chinese software programmers are selected as the group for a subhealth study. In recent years, the subhealth issues of Chinese programmers have received increasing attention from both the patients' side and the medical industry. However, there is not much research on collecting and analyzing the subhealth status data for this group. The present study is dedicated to the programmers as a group, with the collection and analysis of their subhealth status data. The subhealth status of the programmers was investigated using the subhealth scale survey. By using the data-mining algorithm to analyze the collected data, this study attempts to discover the programmers experiencing subhealth status and the factors that affect their health. Therefore, this study lays the foundation for further research on the subhealth of Chinese software programmers.

2. Related work

Among some Chinese subhealth research papers, the study of the designated population mainly focuses on civil servants, teachers, students, drivers or other groups [4,17,24–29]. Ref. [4] compares subhealth distributions based on the basic information of civil servants with the results of subscales and overall scale. Ref. [17] makes subhealth distribution comparison based on teacher's basic information and also conducts a simple response analysis of risk factors and found the most influential factors. Ref. [24] makes comparisons of the subhealth results item by item, according to the basic information and risk factors of 6 provinces and cities. Reference [25] makes comparisons of the subhealth results based on drivers' basic information and risk factors. According to the basic information of Peking University students and subhealth factors in Ref. [26], subhealth results are compared. Reference [27] based on the basic information of college students, conducts a multi-factor regression analysis and regression equations. Reference [28] makes comparisons of the subhealth results based on the basic information of civil servants. Reference [29] is also a subhealth survey of civil servants, teachers, health workers, students, military personnel and migrant workers. The main findings are the comparisons of the distribution of subhealth results under different conditions, which are basically the same as above. Only a small part of regression analysis was done in Refs. [17] and [27], but no subhealth prediction analysis was conducted, and no detailed regression analysis of basic information, risk factors and factors was carried out.

As can be seen from the Table 1, the references [32–34] used logistic regression, boosting algorithm and random forest algorithms, and the accuracy is relatively high. Among them, Reference [32] and [33] adopt the same questionnaire. Different methods show that the boosting algorithm is 3% higher than the logistic regression in terms of prediction accuracy. However, among the above three algorithms, only logistic regression can be drawn the regression equation. Subtracting variables into the regression equation gives quick subhealth judgments that boosting and random forest can not do. However, the subhealth TCM Basic Syndrome Epidemiology Questionnaire [32] was used in Ref. [32], which is also one of the reasons that the accuracy is lower than the results of this test due to the different criteria of the questionnaire and the relatively old time.

The research difficulty in the subhealth field is how to do the classification of different people. The term subhealth describes the human body without any disease being identified by Western medicinal diagnosis, however, these are individuals with a variety of “discomfort” factors and a significant decline in the clinical manifestations of various abilities. The first task in the study of subhealth is to exclude the diseased population, and the remaining population is then further separated into two groups, healthy and subhealthy groups, through the subhealth modeling analysis. As the subhealth status demonstrates itself through more subjective feelings of various symptoms, that is, the “soft index,” it is very suitable for being evaluated through the scale table. Many experts in China are comfortable with the questionnaire method used to evaluate the subhealth status and form the scale table. However, there is not yet a unified, recognized subhealth scale table in China [35–42]. The questionnaires mentioned in Refs. [35–42] have basically the same indicators of measurement. Among them, the Reference [35] is mainly aimed at the physical subhealth survey. Reference [36] mainly provides the research ideas of the questionnaire, but did not put forward specific questionnaire entries. The questionnaire items mentioned in Ref. [39] are mainly the manifestation of TCM symptoms. Reference [40] adds physiological indicators to the questionnaire as a criterion. However, none of the above studies provide a complete scale for reference.

3. Methodology

3.1. Data pre-process

For this study, Chinese programmers were selected as the study sample. A programmer is defined as a person who works in program

Table 1
Compared with recent study.

References	Methodology	Accuracy Rate
Data Mining Research on Subhealth State Pulse Map Feature [30]	Decision tree	72%
	Support Vector Machines	68.3%
	Neural Networks	67.7%
	Bayesian	68.7%
Boosting Algorithm and its Application of the Sub-health Classification Sub-health Influence Factor Model Based on Decision Tree [31]	Decision Tree	54.9%
The Logistic Regression Method Based on the Data Mining Process—The Application of the Sub-health Classification and the Analysis of Effect Factors [32]	Logistic Regression: Backward Method	90.93%
	Logistic Regression: The Best Subset Method	90.54%
Boosting Algorithm and its Application of the Sub-health Classification [33]	Boosting model based on exponential loss function	93.1%
	Boosting model based on Bernoulli distribution loss function	93.1%
Research on Random Forest Based Sub-health State Prediction and Feature Selection Method [34]	Random Forest	91.28%