Correlation-based feature analysis in physical examination indicators can help predict overall underlying health status using machine learning

Haixin Wang
The Key Laboratory for Human Disease Gene Study of Sichuan Province, Chengdu
Ping Shuai
Health Management Center & Physical Examination Center of Sichuan Provincial People's Hospital,Chengdu
Yanhui Deng
The Key Laboratory for Human Disease Gene Study of Sichuan Province, Chengdu
Jiyun Yang
The Key Laboratory for Human Disease Gene Study of Sichuan Province, Chengdu
Yi Shi
The Key Laboratory for Human Disease Gene Study of Sichuan Province, Chengdu
Dongyu Li
Health Management Center & Physical Examination Center of Sichuan Provincial People's Hospital,Chengdu
Tao Yong
Medical Information Center of Sichuan Provincial People's Hospital,Chengdu
Yuping Liu
Health Management Center & Physical Examination Center of Sichuan Provincial People's Hospital,Chengdu
Lulin Huang (huangluling@yeah.net)
The Key Laboratory for Human Disease Gene Study of Sichuan Province, Chengdu

Article

Keywords:

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1535407/v1

Abstract

Because of lacking of the systematic investigation of correlations between the physical examination indicators (PEIs), currently most of them are independently used for disease warning. This results in very limited diagnostic values of general physical examination. Here, we first systematically analyzed the correlations between 221 PEIs in healthy and in 34 unhealthy states in 803,614 peoples in China. Specifically, the study population included 711,928 healthy participants, 51,341 patients with hypertension, 12,878 patients with diabetes, and 34,997 with other unhealthy status. We revealed rich relevant between PEIs in healthy physical status (7,662 significant correlations, 31.5% of all). However, in disease conditions, the PEI correlations changed. We further focused on the difference of these PEIs between healthy and 35 unhealthy physical status, 1,239 significant PEI difference were discovered suggesting as candidate disease markers. Finally, we established machine learning algorithms to predict the health status by using 15%-16% PEIs by feature extraction, which reached 66%-99% precision predictions depending on the physical state. This new encyclopedia of PEI correlation provides rich information to chronic disease diagnosis. Our developed machine learning algorithms will have fundamental impact in practice of general physical examination.

Introduction

The comprehensive primary healthcare system has had a broader impact on human health compared to clinical medical treatment[1]. Health examinations help those who are healthy to improve their understanding of their physical functions and maintain their health status and inform those as to the health benefits conferred by changing unhealthy habits and avoiding risk factors that can lead to disease[2]. Physical examinations can help minimize the distress of diseases[3]. With the population size grows and ages, people's healthcare needs are constantly increasing, and health-care provisions are becoming more sophisticated and in parallel, more costly[4].

Health examinations are common elements of healthcare in developed countries[5]. These checks consist of general blood examination, urine examination, blood glucose examination, blood lipid examination, renal function examination, and so on. However, currently, the physical examination report is assessed mainly based on one or two independent physical examination indicators (PEIs), which can only provide very limited information for physical examiners about their health condition or disease diagnosis[6]. The correlations between PEI in different physical states (i.e. healthy, hypertension, diabetes) have not been systematically investigated, even though they are expected to provide valuable information for public health care, for example by defining a small set of easily measurable PEIs that can be used in the accurate diagnosis of a disease before the disease phenogenesis.

The recent explosion of available health data promises to transform healthcare by improving care quality and as such, improving population health while constraining escalating costs[7]. Health examination centers generate systematic big data that can reveal otherwise undetected underlying health issues[8]. In clinical, there is growing investment in developing big data applications for medical care, such as those

based on artificial intelligence (AI) to diagnose diseases based on clinical images[9]. Although AI can save costs and improve efficiency, especially for early diagnosis and prevention of chronic diseases[10], because of insufficient systematic analysis of PEIs in physical status, currently no prediction models were generated for physical status predictions based on PEIs.

As health-care reform has made impressive progress in the expansion of insurance coverage, now general physical examination industry accumulates big data[11]. By using a large dataset of general health examination of the Chinese population, the present study had three main aims: to determine the correlations among PEIs in healthy and unhealthy (namely, those with underlying chronic disease) patients; to elucidate the relationship between chronic disorders and normal individuals for these PEIs to discovery candidate disease markers; to develop machine learning models that can predict individual health status using a refined set of PEIs. To address these points, we included physical examination data from 80,3614 individuals who visited one health examination center between 2013 and 2018. We included data from 221 PEIs associated with 35 physical conditions, with the majority of unhealthy physical states being due to chronic disease.

Results

Study population

We included 811,244 individuals who attended the Health Management Center & Physical Examination Center between 2013 and 2019. These samples were enrolled from Sichuan province, most of them from Chengdu city. The enrolled samples account for about 1% of the demographics of Sichuan province and 5% of the demographics of Chengdu city. The participants represented 35 healthy states based on either a healthy status or the presence of an underlying disease condition (unhealthy status). Specifically, the study population included 711,928 healthy participants, 46,981 patients with hypertension, 11,745 patients with diabetes, and 32,960 with other unhealthy status (mainly are chronic disease) (Table 1). Besides, 7,630 samples with 12 diseases in replication for prediction were also enrolled in 2019 as a separate dataset. We included 221 PEIs in our analyses, which comprised patient demographic information (age and sex) and life-style indicators (alcohol consumption, tobacco use, etc.).

PEI correlations in participants with a healthy physical status

We first aimed to explore the PEI correlations in healthy status to give a landscape. Among 221 PEIs, we found 7,662 significant correlations (P<0.05/ 24,322 PEI pairs=2×10^{-6}) in all 24,322 PEI pairs correlations (31.5%) (Table 1, Table S1) in those with a healthy physical status (N=711,928, mean age 41.4, female=45.7%). This finding suggests a wide range of correlations between PEIs (Fig. 1). The top 50 correlated PEIs included sex, age, red blood cell count, prealbumin (PAB), history of alcohol intake (alcohol consumption, drinking), alkaline phosphatase level (ALP), tobacco use (smoking) and so on (Fig. 1(a)). Among the 221 PEIs, the number of significantly correlated PEIs also suggested rich correlations

between PEIs (Fig. 1(b)). Of these identified correlations among PEIs in health status, some of them are consistent with the reported literature, but most of them are newly discovered in this study.

General inspection PEIs showed rich relevance to each other or other PEIs. For example, sex showed the richest PEI correlations (151 PEI pairs, males vs. females), including hemoglobin (Hb), creatinine, uric acid (UA), drinking, smoking, body mass index (BMI), etc., which reflect the differences in body shape, physique, and living habits between males and females (Fig. 1, Fig. 2, Table S1). Age also showed strong PEI correlations (125 PEI pairs), such as estimated glomerular filtration rate (eGFB), systolic pressure (SBP), diastolic pressure (DBP), albumin (Alb), and low-density lipoprotein (LDL-C). These findings suggest that with increasing age, body functions systematically change (Fig. 1, Fig. 2, Table S1). We also found 124 PEI correlations with BMI which reflects the strong influence of body shape on PEIs, including UA, high-density lipoprotein (HDL-C), SBP, and DBP (Fig. 1, Fig. 2, Table S1). Blood pressure (BP), which has many physiological meanings, we identified a set of PEIs that correlated with blood pressure (BP), including 125 PEIs for DBP and 124 PEIs for SBP (Fig. 1, Fig. 2, Table S1). Intraocular pressure (IOP) is an important factor for the diagnosis of glaucoma [12]. We found 79 PEIs that were weakly correlated with IOP of the left eye (IOP-L), including IOP of the right eye (IOP-R) SBP, DBP, Alb, BMI, TG, ApoB, drinking, and TC. Similar to IOP-L, 73 PEIs were weakly correlated with IOP-R (Fig. 1, Fig. 2, Table S1).

As expected, blood lipid PEIs display many correlations. For example, 119 PEIs correlated with triglyceride (TG) (Fig. 1, Fig. 3, Table S1). We found 122 PEIs that correlated with HDL-C, with many negative correlations, including TG, UA, and BMI (Fig. 1, Fig. 2, Table S2). The correlation patterns between LDL and HDL showed a specific opposite trend (Fig. 1, Fig. 2, Table S1). Out of expected, living habits have a profound impact on our bodies. Consistently we detected 130 PEIs that correlated with drinkings, such as sex, smoking, Hb, and UA (Fig. 1, Fig. 2, Table S1). Similarly, 128 PEIs were correlated with smoking, including drinking, sex, and age (Fig. 1, Fig. 2, Table S1). We also detected 58 PEIs that weakly correlated with exercise habits (e-habits), including age, eGFB, and SBP (Fig. 1, Fig. 2, Table S1). Tumor marker expression can indicate the occurrence and development of tumors. We detected weak correlations between several tumor markers and PEIs. For example, 88 PEIs were correlated with cytokeratin-19-fragment CYFRA21-1 (CYFRA 21-1); 83 PEIs were correlated with tumor-supplied group factors (TSGF); 64 PEIs were correlated with neuron-specific enolase (NSE); and 64 PEIs were correlated with complexed prostate special antigen (C-PSA) (Fig. 1, Fig. 2, Table S1).

PEI correlations in individuals with an unhealthy physical status

Next, we examined the PEI correlations in 34 unhealthy physical states. In this analysis, we also identified rich correlations in these unhealthy physical states (Table 1). Compared with the healthy physical state, we found fewer significant correlations in PEIs in those with an unhealthy physical status, which might be caused by sample size effect (Table 1, Table S2-S35). Each unhealthy physical state has its only correlation spectrum and most of them are newly discovered in this study. For example, in the hypertension population, we found 4,413 significant correlations in the 221 PEIs of 24,322 PEI pairs (18.3%) (Table S2). The PEI with increased correlations included monocytes (MON) (70 in hypertension vs

six in healthy physical state, the same below), quantitative detection of hepatitis B virus DNA (HBV-DNA) (76 vs 33), quantitative detection of hepatitis C virus RNA (HCV-RNA) (49 vs 8), etc. (Table S2). Those with both hypertension and coronary heart disease (hypertension+coronary) had an increased correlation of RH blood group compared with the healthy cohort (41 vs 9 in normal). Conversely, the numbers of correlations in homocysteine (Hcy) were greatly reduced in unhealthy versus healthy patients (2 vs 120). In diabetes, 10 PEI pairs increased while the remaining 195 PEI pairs decreased; the increased PEIs including MON (41 vs 6), HCV-RNA (42 vs 8), anti-Sc70 (59 vs 31), and HCV-cAg (35 vs 10) (Table S17). These results suggest that under the unhealthy status, the PEIs have changed systematically. Each disease has its own specific PEI spectrum.

Next explored the correlation networks among the PEIs using a qgraph [13], which would show the LinkMode among PEIs. In a healthy status, we found that the PEIs showed rich interactions with both positive and negative directions (Fig. 3). In the unhealthy physical states, each of them showed its unique interaction networks with PEIs (Fig. 4 showed the network of hypertension and diabetes). These results show that there is a dependency relationship between multiple indicators in each physical state, which can be used with the combination in the assessment of physical health.

Candidate PEI markers for unhealthy physical status

To verify and discover new candidate biomarkers or the impact of living habits for disease early diagnosis, we next calculated the difference of each of the 221 PEIs between healthy and unhealthy physical states. In total, we found 1,239 significantly different PEI pairs between healthy and 34 unhealthy physical status (P<0.05/34=0.0014, adjust for 34 unhealthy physical status) (Table 1, Fig. 5, Table S36). For example, 112 PEIs were significantly different between patients with hypertension and healthy people, 100 PEIs were different between hypertension+diabetes and healthy people, and 91 PEIs were different between diabetes and healthy people. Some of them are consistent with previous findings and the rest of them are newly discovered.

For many of the 221 PEI, we detected a difference between healthy and unhealthy physical status, especially in PEIs involved in physique, lifestyles, blood lipids (Fig. 5, Table S36). For example - BMI, we found differences between healthy and unhealthy physical statuses in 16 of the 34 unhealthy physical statuses, including in patients with hypertension (P=0) and gout (P=6.48×10^{-90}). Exercise habits (E-habits) showed 19 differences between healthy and unhealthy status, including in hyperlipidemia (P=1.28×10^{-277}) and diabetes (P=4.20×10^{-29}). Dietary habits also showed differences in 10 unhealthy status, including in chronic pharyngitis (P=2.59×10^{-19}) and cholecystolithiasis (P=9.43×10^{-18}). We detected differences in alcohol intake habits in 20 unhealthy status, including hyperlipidemia (P=0), coronary heart disease (P=4.06×10^{-24}), diabetes (P=1.09×10^{-22}), and Parkinson's syndrome (P=1.43×10^{-17}). We also observed differences in smoking habits in 18 unhealthy status when compared to the unhealthy condition, including in hypertension (P=2.74×10^{-114}), hyperlipidemia (P=2.69×10^{-62}), and Parkinson's syndrome (P=5.12×10^{-29}). We found differences for IOP-R in five unhealthy status compared

with healthy, including in hypertension (P=3.63×10^{-85}) and diabetes (P=2.01×10^{-73}); similar findings were produced for IOP-L (Fig. 5, Table S36). For lipids PEIs, we also observed differences between 34 unhealthy and healthy status. For example, LDL-C was detected in 21 unhealthy status, including hypertension (P=0) and diabetes (P=2.95×10^{-212}). HDL-C was detected in 17 unhealthy status, including in diabetes (P=1.92×10^{-177}) (Fig. 5, Table S36). We further conducted a detailed analysis of HDL-C and diabetes and found those with low HDL-C showed a significantly higher risk of developing diabetes than those with average values (1.26-1.75 mmol/L) in this population. Of note, those with high HDL-C levels also showed an elevated risk of developing diabetes (Fig. 6).

Tumor-associated antigens also display significant differences between healthy and unhealthy status. For example, CYFRA 21-1 was detected in 10 unhealthy status, including hypertension+diabetes (P=3.71×10$^{-97)}$ and diabetes (P=4.52×10^{-70}). CEA1 was detected in 12 unhealthy status, including hypertension+coronary (P=9.59×10^{-29}) and diabetes (P=1.73×10^{-18}). Alpha-fetoprotein (AFP) was detected in hepatopathy (P=1.08×10^{-28}). C-PSA was detected in hypertension+coronary (P=8.38×10^{-20}). Finally, the carbohydrate antigen CA724 (CA 72-4) was detected in asthma (P=9.92×10^{-13}), gout (P=3.53×10^{-7}), and coronary+diabetes (P=4.06×10^{-5}) (Fig. 5, Table S36). Among other PEIs, we also detected significant differences between healthy and unhealthy status. For example, we found differences in urine sugar levels (U-GLU) in nine unhealthy status, including in diabetes and its associated diseases. The eosinophil rate (eo%), was found in five unhealthy status, including asthma (P=1.38×10^{-129}) and rhinallergosis (P=4.05×10^{-18}). Whole blood iron levels (WB-Fe) was found in 11 unhealthy status, including hypertension (P=2.52×10^{-69}). We detected PH in 11 unhealthy status, including diabetes (P=1.97×10^{-239}), hypertension (P=2.41×10^{-166}), hypertension+diabetes (P=9.90×10^{-32}), and gout (P=9.82×10^{-15}). We found potassium (K+) in five unhealthy status, including hypertension (P=1.98×10^{-119}) and hepatitis B (P=3.13×10^{-10}). We also detected differences in magnesium (Mg2+) in hypertension+diabetes (P=3.14×10^{-58}) and diabetes (P=5.10×10^{-52}). Hcy (an indicator of cardiovascular disease) was detected in eight unhealthy status, including hypertension (P=1.97×10^{-136}) and Parkinson's syndrome (P=1.76×10^{-7}) (Fig. 5, Table S36). These results provide a set of candidate markers for chronic diseases early diagnosis.

Machine learning to predict healthy and unhealthy physical status from PEIs

A key objective of this study was to apply PEI data and machine learning technology to develop algorithms that can predict a common disease based on general physical examination. We tried three machine learning models, including kernelized support vector machine (SVM), multilayer perceptron (MLP), and random forests. MLP prediction models only resulted in a low f1_score, recall, and precision in our initial training data. It takes tens of hours for the SVM model to do a binary classification, so we excluded MLP prediction models and SVM prediction models for further training. We found that random forest is more suitable for our data. It only takes 2–3 minutes to do a binary classification, and the prediction effect of random forest is much better than that of MLP and SVM. However, the random forest

could not give good performance in the multi-class classification of all the physical status. Finally, we tried to use binary classification to classify each pair of healthy and unhealthy physical status (e.g. hypertension and healthy people; Parkinson's syndrome and healthy people) and we obtained relatively better performance than the multi-class classification. Then we tried to optimize this prediction algorithm. Because the data were characterized by serious category imbalance, a random under-sampling method was adopted that balances the data by randomly selecting the data subset of the target class. In each physical status, the top 15% or 16% representative PEIs were extracted for prediction by feature extraction. The advantage of this method is that it is usually very fast and completely independent of the model applied after feature selection.

Finally, in the random forests algorithm prediction of each pair of healthy and unhealthy physical status, the area under the curve (AUC) of receiver operating characteristic curve reached 66%~99% depending on the unhealthy physical status (average 87.6%) (Fig. 7, Table 2 and Table S37 and 38). For classification, AUC values more than 90% indicated excellent performance, and values from 80% to 90% indicated good performance. Our algorithm provided high-precision predictions in 18 of the 34 unhealthy physical status (AUC>90%), good performance for another 9 of the unhealthy physical status (90% >AUC>80%). In our algorithm, patients with heart-related diseases showed excellent performance. For example, by extraction 30 PEI features (age, leukocyte count, monocytes, Mon%, mean corpuscular volume, red blood cell count, red cell distribution width, lymphocyte rate, platelet count, low-density lipoprotein, high-density lipoprotein, total cholesterol, carcinoembryonic antigen 1, albumin, albumin-globulin, cystatin c, glucose, urine sugar, urine creatinine, estimated glomerular filtration rate, creatinine, urea, waistline, waist-hip Ratio, body mass index, operation history, systolic pressure, height, neck size, and anamnesis), Hypertensive+Diabetes+Coronary Heart Disease provides 99% AUC just using 909 training samples and 387 validation samples (f1-score (95%CI), 0.96(0.95-0.96); accuracy (95%CI): 0.95(0.94-0.97); specificity (95%CI): 0.95(0.94-0.95); recall (sensitivity) (95%CI): 0.95(0.94-0.97). In our algorithm, patients with Parkinson's syndrome provides 97% AUC using 192 training samples and 83 validation samples (f1-score (95%CI), 0.91(0.90-0.91); accuracy (95%CI): 0.90(0.89-0.90); specificity (95%CI): 0.87(0.79-0.94); recall (95%CI): 0.90(0.89-0.91). For hepatic adipose infiltration, our algorithm also provided good prediction performance using 803 training samples and 115 validation samples (f1-score (95%CI), 0.82(0.78-0.87); accuracy (95%CI): 0.81(0.76-0.86) ; specificity (95% CI): 0.75(0.67-0.82); recall (95% CI): 0.82(0.77-0.87) and AUC (95% CI): 0.92(0.89-0.94). For chronic rhinitis, we got the lowest prediction performance in this study (AUC (95%CI):0.66(0.60-0.72)). When all unhealthy physical status were classified as one “unhealthy” status together, our algorithm also provided good predictions: f1-score (95%CI): 0.83 (0.83-0.83); accuracy (95%CI): 0.82 (0.82-0.82); specificity (95%CI): 0.81(0.81-0.81); sensitivity (95%CI): 0.84 (0.84-0.84) and AUC (95%CI): 0.9 (0.90-0.90). These results suggested that by using feature extraction of the PEIs (15-16% of all 221 PEIs) just by using a small number of samples, our random forest algorithms provided good performance for majority unhealthy physical status predictions.

To further validate our random forest algorithm prediction model, we did a replication analysis of 12 diseases in another new dataset. The results are presented in Fig. 8 and Table 3. The ROC of the replication data achieved 0.63–0.98 (average 0.90) (Fig. 8), suggesting a good performance of the

prediction effect, based on the limited samples. For the rest of the diseases, we did not obtain enough samples in the new dataset for replication (<100 samples).

In this study, the top 15% or 16% representative PEIs were extracted for random forest prediction by feature extraction in each physical status (Table S38), which reached 66%–99% precision in predictions, depending on the physical state. In total, 161 PEIs were used for the random forest prediction of 35 pairs of health statuses. Some PEIs were used more frequently than others, suggesting their important physiological values for the human body. The top 20 used PEIs included monocyte counts (36 health statuses used, the same as below, 100%), anamnesis (33, 92%), age (32, 98%), albumin (31, 86%), estimated glomerular filtration rate (30, 83%), systolic pressure (27, 75%), waistline (27, 75%), red cell distribution width (26, 72%), creatinine (23, 64%), neck size (23, 64%), operation history (23, 64%), red blood cell count (23, 64%), urea (22, 61%), waist-hip ratio (22, 61%), BMI (21, 61%), gender (21, 61%), height (20, 56%), glucose (19, 53%), hemoglobin (19, 53%) and platelet count (19, 53%). Some PEIs were rarely used, suggesting their unique indication of a certain disease. For example, sodium was only selected for cholecystolithiasis prediction, and cholinesterases were only selected for rhinallergosis prediction. Our results provide proof for predicting health conditions just using a set of PEIs.

Discussion

This study has produced correlation maps of 221 routine PEIs using physical examination data obtained from a Chinese population of 811,244 individuals of 35 healthy or unhealthy physical status (mainly chronic diseases). We detected a large number of correlations among PEIs in healthy or unhealthy physical states; furthermore, these correlations differed according to the 34 unhealthy physical conditions analyzed. Most of the correlations are newly observed in this study. We found that a wide range of correlations among PEIs, such as sex, age, BMI, blood lipids, blood pressures, cancer-related indicators, lifestyles including drinking, smoking, e-habits. Improving our understanding of these PEI interactions will help explain disease mechanisms and pathogenesis. Our results fill the gap of systematic PEI analysis and provide rich information about how PEIs might reflect underlying health conditions. These findings provide rich information to further improve healthcare researches and clinical practice.

One of the unexpected findings from our analysis was that patients with hypertension showed more correlations between HBV-DNA and HCV-RNA to other PEIs than a healthy cohort. Similarly, we found a strong correlation between hepatitis C virus and other PEIs in diabetes, suggesting that patients infected with hepatitis C may be more susceptible to diabetes. This finding implicates a phenomenon whereby viral infection can make an individual more susceptible to developing a chronic disease. For these people, antiviral therapy might be taken into consideration while treating hypertension and diabetes.

Biomarker discovery and development for clinical research, diagnostics and therapy monitoring in clinical trials are key areas of medicine and healthcare [14]. In this study, we presented many candidate markers for chronic disease. For example, we found that IOP indicators, which are considered to be a relatively independent marker for glaucoma [15], are closely associated with hypertension, diabetes, and

hypertension with diabetes. These results suggest that IOP might be affected, to some extent, by systemic diseases and might be used as one of the clinical markers of these diseases early diagnosis. Our results confirmed that low HDL-C level is a risk factor for diabetes, especially in women[16]. This result suggests that improving HDL-C levels through dietary supplementation might be an effective way to prevent diabetes in patients with low HDL-C levels. However, based on our results, excessive HDL-C supplementation is also a risk factor; therefore, HDL-C supplementation should aim to bring HDL-C levels within a normal range[17]. We detected a significant increase in AFP in hepatopathy when comparing a healthy cohort, which confirms AFP increase is an increased risk factor for primary liver cancer in hepatopathy[18]. K^+ has significant effects on hypertension[19] and Cl^-, and Mg^{2+} has significant effects on diabetes, suggesting that modulation of these ions might have effects on these conditions. Living habits, such as exercise, smoking, and drinking, have a more profound impact on the body than we had expected. For example, exercise, drinking, or smoking history have a strong impact on hyperlipidemia[20], as evidenced by comparision to healthy status. This finding suggests that by adjusting these living habits, hyperlipidemia should improve.

Because the current physical examination conclusion is generally based on a relatively independent single or several prior indicators to advise on the results of physical examination, many of the results given are ambiguous, and the value of judging the health status of the examinees is very limited [21]. There is an urgent need for a more accurate index system and method to judge the health status of physical examinees. In the final part of our study, we developed random forest machine learning algorithms that can predict diseases through 15%-16% of all 221 PEIs with good performance of prediction (AUC:66%~99%; average 86%). For each disease, we defined about 30 contributed PEIs by feature extraction. In most of our prediction algorithms, only a few hundreds of samples were needed to give good prediction performance for many chronic diseases. This finding suggests machine learning on PEI data can be used to help predict the true condition of the examiners, identify “at-risk” patients, and indicate the most relevant follow-up physical examinations for affected individuals. However, the operation of this method is not as convenient as that of a multiclassification. We have tried multiclassification models before, unfortunately, we have not achieved good prediction results. When using multiple binary classification models, the total number of times to run the models is more, but the time required to run a complete model is very short, and the prediction effect using binary classification is also better.

This study also has some limitations. Firstly, missing data are a common problem in all types of research. Overall, mean imputation may lead to inefficient analyses, and it commonly produces biased estimates of the association(s) investigated. Secondly, the data used in this study were taken from a single institution, which may introduce significant bias and limitations in ML generalization into real-world applications, especially with discrepancies in EMRs and patient populations. Thirdly, some models had very small datasets; for example, Parkinson's model included ~ 250 patients. Small datasets may cause insufficient power for MI models. Fourthly, a multiclassification model may support a stronger case for AI and ML to be used as a clinical tool to improve our decision making than multiple binary

classification models. However, we did not obtain good prediction results by multiclassification in this study.

In summary, we systematically explored the correlation between various PEIs and their relationship with chronic diseases and established machine learning prediction models to predict health status. This study provides abundant information to better understand the physiological and pathological characteristics of the human body as a system. Importantly, we have identified modifiable factors and directions for disease prediction, diagnosis, and treatment. Our developed machine learning algorithms can be immediately applied to clinical practice to assist in the judgment of physical examination results.

Methods

Study approval

This study is a general retrospective study. The study was approved by the institutional ethics committee of Sichuan Provincial People's Hospital (No.2019276) according to the following rules: 1) Ethical review of biomedical research involving human beings (Order of the state health and Family Planning Commission of the people's Republic of China, 2016); 2) WMA Declaration of Helsinki (2013) and 3) CIOMS International Ethical Guidelines for Biomedical Research Involving Human Subjects. This study just extracts the clinical PEI data; no patient's identity was involved in this study. The researchers try our best to protect the information provided by patients from disclosing personal privacy and hereby applied for exemption from informed consent.

Study Participants

PEI data were obtained from 811,244 Han Chinese patients visiting the Health Management Center & Physical Examination Center of Sichuan Provincial People's Hospital between 2013 and 2018. The total cohort captured participants with 35 different reported health conditions, including 711,928 reported healthy participants and 91,686 unhealthy participants. The unhealthy cohort included 46,981 patients with hypertension, 11,745 with diabetes, and 32,960 with another unhealthy status (Table 1).

Detected PEIs

Only the PEIs that were recorded by the same methods were included in this study. In total, 229 PEIs were initially collected: eight PEIs that were detected in a few individuals were excluded, leaving 221 PEIs for further analysis (**Table** 1). This PEIs included the levels of biochemical indicators and the results of blood tests. Patient lifestyles and disease conditions were also investigated during the physical examination.

Data processing

The PEIs with string variables were converted to integer variables for data analysis. Categorized variables were digitally coded for further calculation. The mean value imputation method was used for missing

data (about 20% of all data). For individuals who participated in more than one physical check, the average values of each PEI were used for data analysis.

Statistical Analyses

The Pearson correlation coefficient (PCC) method was used to calculate the correlations between two PEIs (for example, x and y) in R; this method measures the linear dependence between two variables. PCC correlation (r) (1) and *P* values (2) were calculated using the following formulae [22]:

(1)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

(2) $P = 1 - F.DIST(((n-2)*r^2)/(1-r^2), 1, n-2)$

$df = n - 2$

n = number of x-y data pairs

Total sample size required when using the correlation coefficient (r), when two-sieded α=0.05, β=0.20. If r=0.05, we need 3,134 samples; if r=0.10, we need 782 samples; if r=0.25, we need 123 samples; if r=0.5, we need 29 samples. The general formula for the correlation sample calculating is listed as the following (3) [23]:

(3) *r* = expected correlation coefficient

$C = 0.5 \times \ln [(1 + r)/(1 - r)]$

N = Total number of subjects required

Then

$N = [(Z_\alpha + Z_\beta) \div C]^2 + 3.$

A linear regression model (lm) was used to compare PEIs between the reported healthy status and unhealthy status adjusted for sex and age in the R package[23]. The odds ratio of HDL-C level was calculated by using generalized linear models (glm) and adjusted for age in the R package[24]. The 5% CI

can be calculated by using model = fit_glm. The correlation interaction network was conducted using qgraph[25].

Machine learning

Three machine learning models, including kernelized support vector machine (SVM)[22,26], multilayer perceptron (MLP)[23,27-29] and random forest[30] were tested to get the prediction performance of the PEIs. By using MLP algorithm prediction in the neural network to predict health and each of the 34 unhealthy status (multi-classification), it could not achieve good results. We further tried prediction the healthy from each unhealthy statues by the binary classification method, the F1 value of the prediction each result is very close to zero. By using SVM algorithm prediction for making a multi-classification prediction, the highest F1_socre of cholecystolithiasis 0.70, but that of most other types of diseases is 0.00. We also tried the binary classification method, but all the results were relatively poor. When a random forest algorithm is used for prediction for multi-classification (health and each of the 34 unhealthy status), the F1 value of health status can reach 0.80-0.90, but the F1 value of unhealthy status is about 0.00-0.40. Then, we further have chosen a forest algorithm and optimized the random forest algorithm. First, due to the uneven distribution of the sample numbers of healthy and non-healthy status, and the law of large numbers[30], we used downsampling strategy for sample randomly used. Because the data were characterized by serious category imbalance, a random under-sampling method was adopted that balances the data by randomly selecting the data subset of the target class. Second, we used PEI feature extraction strategy to extract the most contributed PEI for each healthy and unhealthy status. Feature extraction adopts the strategy of univariate statistics in automatic feature selection. Univariate statistics select features with high confidence according to the statistical significance of the relationship between each feature and the target. This process can be achieved by using feature_selection in scikit-learn. Finally, in each healthy and non-healthy status, the top 15% or 16% representative PEIs were extracted for prediction by feature extraction. The advantage of this method is that it is usually very fast and completely independent of the model applied after feature selection. Then, the data were randomly divided such that 30% constituted the test set, and the remaining 70% were randomly divided again, with 70% as the training set for the training model and 30% as the validation set for the evaluation model. In the process of improving the generalization performance of the model by adjusting parameters, a cross-validation method with a grid search was adopted, which can be implemented by GridSearchCV provided by scikit-learn (Table S37 and Supplementary code). In our random forest model, in the randomforestclassifier() function, criterion = 'entropy', random_ State = 3, and for the random forest model of binary data, we mainly adjust n_ estimators, max_ depth and min_ samples_ leaf; these three parameters make the model achieve better results, while the other parameters are defaults. The evaluation of the model effect is mainly based on the use of sklearn. metrics F1 in_ score and ROC_Curve (AUC is calculated according to true positive rate and false-positive rate, and the ROC curve is drawn accordingly).

Declarations

Data availability

All the data can be found in the supplementary datasets.

Code availability

The R script used in this research is publicly available and can be found in the supplementary data.

Ethics approval and consent to participate

The study was approved by the institutional ethics committee of Sichuan Provincial People's Hospital (No.2019276) according to the following rules: 1) Ethical review of biomedical research involving human beings (Order of the state health and Family Planning Commission of the people's Republic of China, 2016); 2) WMA Declaration of Helsinki (2013) and 3) CIOMS International Ethical Guidelines for Biomedical Research Involving Human Subjects. This study just extracts the clinical PEI data; no patient's identity was involved in this study. The researchers try our best to protect the information provided by patients from disclosing personal privacy and hereby applied for exemption from informed consent.

Consent to publish

The authors agree on the publication.

Competing interests

The authors declare no competing interests related to this paper.

Funding

This research project was supported by the National Natural Science Foundation of China (81970839, 81670895, 82121003 and 81870683; the Department of Science and Technology of Sichuan Province, China (2021YFS0033, 2017JZ0039 and 2020JDTD0028; the Grant from Chinese Academy of Medical Sciences (No. 2019-I2M-5-032).

Authors' Contributions

L.H. designed the study. L.Y, P.S., J.Y., Y.S., D.L., and T.Y. enrolled all the participants. L.H., H.W., and P.S. performed the data analysis. Y.D. did the machine learning prediction models. L.H. wrote the manuscript. All of the authors critically revised and provided final approval for this manuscript.

References

1. Steenhuis, S., Groeneweg, N., Koolman, X. & Portrait, F. Good, better, best? A comprehensive comparison of healthcare providers' performance: An application to physiotherapy practices in primary care. Health Policy **121**, 1225–1232, doi:10.1016/j.healthpol.2017.09.021 (2017).

2. Liu, Q., Tian, X., Tian, J. & Zhang, X. Evaluation of the effects of comprehensive reform on primary healthcare institutions in Anhui Province. BMC Health Serv Res **14**, 268, doi:10.1186/1472-6963-14-268 (2014).
3. Perry, H. B., Shanklin, D. S. & Schroeder, D. G. Impact of a community-based comprehensive primary healthcare programme on infant and child mortality in Bolivia. J Health Popul Nutr **21**, 383–395 (2003).
4. Lennox, N. G., Green, M., Diggens, J. & Ugoni, A. Audit and comprehensive health assessment programme in the primary healthcare of adults with intellectual disability: a pilot study. J Intellect Disabil Res **45**, 226–232, doi:10.1046/j.1365-2788.2001.00303.x (2001).
5. Dodd, R. *et al.* Organisation of primary health care systems in low- and middle-income countries: review of evidence on what works and why in the Asia-Pacific region. BMJ Glob Health **4**, e001487, doi:10.1136/bmjgh-2019-001487 (2019).
6. Keenan, G. M. Big Data in Health Care: An Urgent Mandate to CHANGE Nursing EHRs! Online J Nurs Inform **18** (2014).
7. van Ginneken, E. Perennial Health Care Reform–The Long Dutch Quest for Cost Control and Quality Improvement. N Engl J Med **373**, 885–889, doi:10.1056/NEJMp1410422 (2015).
8. Krogsboll, L. T., Jorgensen, K. J., Gronhoj Larsen, C. & Gotzsche, P. C. General health checks in adults for reducing morbidity and mortality from disease. Cochrane Database Syst Rev **10**, CD009009, doi:10.1002/14651858.CD009009.pub2 (2012).
9. Frieden, T. R. SHATTUCK LECTURE: The Future of Public Health. N Engl J Med **373**, 1748–1754, doi:10.1056/NEJMsa1511248 (2015).
10. Goroll, A. H. Toward Trusting Therapeutic Relationships–In Favor of the Annual Physical. N Engl J Med **373**, 1487–1489, doi:10.1056/NEJMp1508270 (2015).
11. Krogsboll, L. T., Jorgensen, K. J. & Gotzsche, P. C. General health checks in adults for reducing morbidity and mortality from disease. Cochrane Database Syst Rev **1**, CD009009, doi:10.1002/14651858.CD009009.pub3 (2019).
12. Vigilante, K., Escaravage, S. & McConnell, M. Big Data and the Intelligence Community - Lessons for Health Care. N Engl J Med **380**, 1888–1890, doi:10.1056/NEJMp1815418 (2019).
13. Furr, R. M., Fleeson, W., Anderson, M. & Arnold, E. M. On the Contributions of a Network Approach to Personality Theory and Research. Eur J Pers **26**, 437–439, doi:10.1002/per.1871 (2012).
14. Obermeyer, Z. & Emanuel, E. J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med **375**, 1216–1219, doi:10.1056/NEJMp1606181 (2016).
15. Sultan, M. B., Mansberger, S. L. & Lee, P. P. Understanding the importance of IOP variables in glaucoma: a systematic review. Surv Ophthalmol **54**, 643–662, doi:10.1016/j.survophthal.2009.05.001 (2009).
16. Aryal, M. *et al.* Evaluation of non-HDL-c and total cholesterol: HDL-c ratio as cumulative marker of cardiovascular risk in diabetes mellitus. Kathmandu Univ Med J (KUMJ) **8**, 398–404, doi:10.3126/kumj.v8i4.6239 (2010).

17. Borggreve, S. E., De Vries, R. & Dullaart, R. P. Alterations in high-density lipoprotein metabolism and reverse cholesterol transport in insulin resistance and type 2 diabetes mellitus: role of lipolytic enzymes, lecithin:cholesterol acyltransferase and lipid transfer proteins. Eur J Clin Invest **33**, 1051–1069, doi:10.1111/j.1365-2362.2003.01263.x (2003).
18. Diness, J. G. *et al.* Effects on atrial fibrillation in aged hypertensive rats by Ca(2+)-activated K(+) channel inhibition. Hypertension **57**, 1129–1135, doi:HYPERTENSIONAHA.111.170613 [pii] 10.1161/HYPERTENSIONAHA.111.170613 (2011).
19. Chanoine, P. & Spector, N. D. Hyperlipidemia, eating disorders, and smoking cessation. Curr Opin Pediatr **20**, 734–739, doi:10.1097/MOP.0b013e32831a6bed (2008).
20. Saito, Y. Secondary hyperlipidemia due to obesity and alcohol drinking. Nihon Naika Gakkai Zasshi **81**, 1784–1787 (1992).
21. Mehrotra, A. & Prochazka, A. Improving Value in Health Care–Against the Annual Physical. N Engl J Med **373**, 1485–1487, doi:10.1056/NEJMp1507485 (2015).
22. Brown, B. W., Jr., Lucero, R. J. & Foss, A. B. A situation where the Pearson correlation coefficient leads to erroneous assessment of reliability. J Clin Psychol **18**, 95–97, doi:10.1002/1097-4679(196201)18:1<95::aid-jclp2270180131>3.0.co;2-2 (1962).
23. Kew, M. Alpha-fetoprotein in primary liver cancer and other diseases. Gut **15**, 814–821, doi:10.1136/gut.15.10.814 (1974).
24. McCuish, E., Bouchard, M., Beauregard, E. & Corrado, R. A Network Approach to Understanding the Structure of Core Symptoms of Psychopathic Personality Disturbance in Adolescent Offenders. J Abnorm Child Psychol **47**, 1467–1482, doi:10.1007/s10802-019-00530-9 (2019).
25. Lima, M. P., Machado, W. L. & Irigaray, T. Q. Predictive factors of treatment adherence in cancer outpatients. Psychooncology **27**, 2823–2828, doi:10.1002/pon.4897 (2018).
26. Haldar, P. *et al.* Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med **178**, 218–224, doi:10.1164/rccm.200711-1754OC (2008).
27. Colling, J. Designing clinical research studies: Part I. Urol Nurs **23**, 357–360 (2003).
28. Colling, J. Designing clinical research studies: Part II. Urol Nurs **23**, 449–451, 448 (2003).
29. Colling, J. Designing clinical research studies: Part III. Urol Nurs **24**, 58–61 (2004).
30. Laurikkala, J. P., Kentala, E. L., Juhola, M. & Pyvkko, I. V. A novel machine learning program applied to discover otological diagnoses. Scand Audiol Suppl, 100–102, doi:10.1080/010503901300007218 (2001).

Figures

Figure 1

The PEI correlations were detected in the healthy cohort. (a) A correlation map of the top 50 correlated PEIs, each of which had >114 significant correlations with other PEIs (FDR<0.05). (b) The number of statistically significant correlations detected in the healthy population of each PEI.

Figure 2

The correlation directions of typical PEIs in healthy physical conditions. The r values were calculated by the PCC method. See **Table** 1 for detailed PEI information.

Figure 3

PEI networks in healthy physical status. In the weighted graphs, the green edges indicate positive weights, and the red edges indicate negative weights. The color saturation and the width of the edges correspond to the absolute weight and scale relative to the strongest weight in the graph, respectively. The circular layout shows how well the data conforms to the model while the force-oriented layout shows how each node (connected and unconnected) repulses the other, and how connected nodes attract each other. See also Supplementary Fig.s.

Figure 4

PEI networks in hypertension (a) and diabetes (b). In weighted graphs, green edges indicate positive weights, and red edges indicate negative weights. The color saturation and the width of the edges correspond to the absolute weight and scale relative to the strongest weight in the graph. At a minimum, the edge with absolute weight at this value is omitted. The circular layout is convenient to see how well the data conform to a model but to show how the data clusters, another layout is more appropriate. A force-oriented layout was created by specifying layout = "spring". In principle, what this function does is that each node (connected and unconnected) repulses each other, and connected nodes also attract each other. The full view of these Fig.s is provided in Supplementary Fig.s.

Figure 5

Representative candidate markers for unhealthy physical status. A linear regression model was used to compare PEIs between healthy physical states and unhealthy physical states after adjusting for sex and age. Significantly different PEIs (P <0.05) after Bonferroni adjustment (P<0.05/34 unhealthy states=1.4×10^{-3}) are shown. See also Table S36.

Figure 6

Odds ratios for HDL-C concentration in plasma from those with a normal physical status and those with diabetes. Both male and female subjects were included in this study.

Figure 7

Machine-learning prediction of the 35 physical status by the random forest algorism. The receiver operating characteristic curve takes the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The horizontal axis represents the proportion of the actual negative instances in the positive class predicted by the classifier to all the negative instances, while the vertical axis represents the proportion of the actual positive instances in the positive class predicted by the classifier to all the positive instances. The AUC is the area under the ROC curve.

Figure 8

Machine-learning replication of the prediction of the 12 physical status by the random forest algorism. The receiver operating characteristic curve takes the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The horizontal axis represents the proportion of the actual negative instances in the positive class predicted by the classifier to all the negative instances, while the vertical axis represents the proportion of the actual positive instances in the positive class predicted by the classifier to all the positive instances. The AUC is the area under the ROC curve.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WangSupplementaryMaterials.pdf