Health Checkup could Reveal Chronic Disorders with Support from Artificial Intelligence

Shuangquan Li[1,2],Tongbin Zhang[1,2],Chuandi Pan[1,2*],Li Cai[3]

[1]*School of 1st Clinical Medical Sciences, School of Information and Engineering, Wenzhou Medical University, Zhejiang, China*

[2]*Department of Computer Technology and Information Management, The First Affiliated Hospital of Wenzhou Medical University, Zhejiang, China*

[3]*Distance Education College, Shaanxi Normal University, Xi'an 710062, Shanxi, China*

ABSTRACT

After decades of practice, healthcare specialists have not reached the conclusion: to what extent health checkup could improve the quality of care. In this paper, we join this debate with a larger health checkup cohort than most of the previous studies. In addition, we examine the health checkup potential in a new task: identifying chronic diseases for the individual with the support of digital health (Health IT) and Artificial Intelligence (AI). Our results show that with the assistance of Health IT and AI, the health checkup data could identify many types of chronic disorder with high precision. In addition, we found specific associations between occurrence of chronic disease and results of lab tests in the health checkup. Using these associations not only improves the predictive performance but also points out that the health checkup could have a role in preventive care. Furthermore, the results provide some evidence for the healthcare organizations to design a better and cheaper checkup service, which uses a smaller number of tests but preserves a similar predictive capacity. Therefore, the health checkup, with support from Health IT and AI, has rich potential to improve the quality of care in predictive and preventive tasks.

Keywords: *Health checkup, Artificial Intelligence, Machine Learning, Health IT, Chronic disease prediction*

I. INTRODUCTION

The clinical benefits of health checkup (also called general medical examination) practice are still being debated [1,2]. In the other hands, when focusing on specific diseases and the screening role of the checkup practice, the checkup practice show clear evidence in improving the quality of care and patient's satisfaction [3]. The diverse positive/non-positive conclusions on the impact of health checkup also associate with the prevalence of adopting this practice. For example, among the studies lists above, the non-positive conclusions for the checkup practice come mostly from the UK and some European countries, where the practice remains unpopular with the general populace [2] or the practice is not widely-standardized and routine. Meanwhile, many positive conclusions for checkup come from Japan and South Korea, where the annual checkup is nationally standardized or required by the state department of labor [3,4]. In China, the clinical benefit of health checkup is still an open question. Medical research in China mostly uses the checkup as a data source to conduct descriptive studies on specific cohorts [5,6]. Addressing the clinical impact of this practice in improving the quality of care in China has not been conducted thoroughly.

The inconclusiveness on the clinical impact of health checkup may largely due to the fact that the health checkup data content has been under-utilized. To conclude the impact of the checkup practice, these above studies only applied basic medical statistics [7]. In addition, each study above generally included less than 5000 patients, which is significantly less than the amount of data available in one caretaker today. In addition, we have not seen many applications of artificial intelligence / machine learning (AI/ML) techniques, which already had some initial successes in general health data [8-11], in utilizing the big health checkup data to improve the quality of care. There are several challenges in electronic health record (EHR), including electronic/digital health checkup data, limiting the success in applying AI/ML techniques, including noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases [12,13]. If we can overcome these challenges, the large electronic health checkup may show unexpectedly large potential in disease-predictive tasks, which is similar to what other studied in general EHR have showed [10,14].

In this work, we demonstrate the potential impact of the checkup in identifying chronic diseases using a larger cohort (17,000 subjects) than many other prior works. Using statistical feature selection techniques in AI/ML [15], we reveal specific associations among occurrence of chronic diseases and results of medical lab tests in the health checkup, in which the lab tests are not used officially in specific disease diagnosis. Using these associations not only improves the predictive performance but also points out that the health checkup could have a role in preventive care. We conduct the research by integrating the annual checkup and the outpatient medical record data at the First Affiliated Hospital (1AH), Wenzhou Medical University, Zhejiang, China. Although the study only involves one healthcare provider (1AH), the provider is among the 20-largest hospitals in China [16]. The caretaker fully implements the Chinese national standard for the health checkup and electronic medical record. Therefore, the conclusions in this work could be very likely repeated in other major healthcare providers in China.

II. METHODS

A. The health checkup protocol in the 1AH

The 1AH has an independent health checkup department providing the checkup services for the general population in Wenzhou city, Zhejiang China and all patients at 1AH. The checkup department serves about 300,000-400,000 cases per year. The department offers 18 different checkup packages, in which 4 packages are for general checkup purposes. More details about these packages, including labtests covered in each package, could be found at http://oa.wzhospital.cn:8030/tjzx/Tcxz.aspx. Among these 18 packages, there is a core package, which costs $100 without insurance support ($10 with general insurance support), fully implementing the Chinese national standard for the checkup. The core package contains 97 lab tests, as shown in the supplemental table 1. The other packages also cover these 97 tests.

B. Acquire and preprocess data

In this study, we acquired the outpatient dataset from 1AH for chronic disease diagnosis information and query these checkup data from the health checkup department for these patients. Among the data sectors at the 1AH (checkup, outpatient, inpatient, and non-hospitalization public service), the outpatient contains the highest number of chronic-disease patients with multiple follow-up visits for further validation. In this work, the chronic disease outpatient HER was collected between October 2010 and August 2014, specified by the research sponsor. The dataset contains information on 16,310 patients with chronic diseases (identified by ICD code version 10 [17]). There are 73 unique ICD codes for chronic diseases; however, one disease may have multiple ICD codes. By manual checking, we found that the dataset covers 29 different chronic diseases. We completely removed the patients' demographic information according to the patient privacy regulation in China and the requirements of the research sponsor. These patients made 265,903 visits (identified by visit number) between 2010 and 2014 (averagely 16 visits per patient). We show the number of visits per patient distribution in figure 1a. Figure 1b shows the distribution of comorbidity size per patient. Among these, 1,919 patients only had one visit; therefore, we do not use these patients' information in the analysis. 9,746 patients only had one chronic disease; meanwhile, 6,564 patients showed comorbidity among at least two diseases.

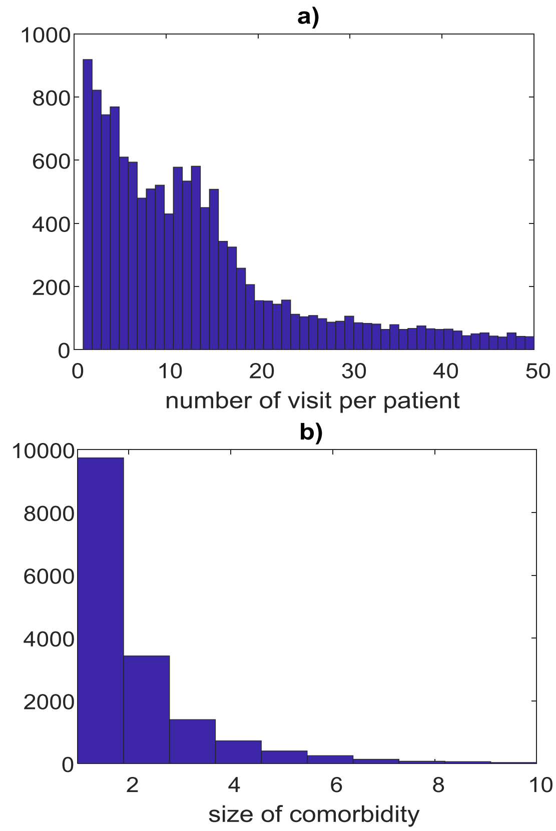

Fig.1 Distribution of number of visit (a) and comorbidity size (b) for each patient

In addition, to form the control set for the statistical analysis, we acquired the checkup from random 1000 subjects who show no abnormality between 2010 and 2014. These subjects made 1125 visits. These subjects had neither inpatient nor outpatient visits at 1AH. Therefore, by the scope of the project, we may assume that they are healthy subjects. We chose the control class subject such that their checkup visits are uniformly distributed between 2010 and 2014 and the subjects' ages are uniformly distributed from 20 to 50. We only limited the control set to 1000 subjects since most of the specific chronic diseases set has less than 1000 patients.

When linking the data from the outpatient sector and the checkup department, we removed the tests which do not belong to the checkup core package (containing 97 tests). The removed tests are either rare/expensive (more than $30) or too specific for disease diagnosis (not for checkup purpose). Therefore, using these tests would limit the predictive capacity of the checkup data. We manually translated the test names from Chinese into English and re-identify these tests because some tests have multiple test ID at 1AH.

C. Identify and validate the occurring diseases-lab test results associations

For the validation purpose, for each disease (positive class), we divided the dataset into two the training set and test set, as shown in figure 2. The training set only contains patients having discovery date, or the earliest date when the patient was diagnosed with the disease, prior to January 1, 2014; while the test set only contains patients having discovery date after January 1, 2014. Then, for each disease analysis, we setup the feature table as follow. In the feature table, each patient represents a row in the table; while each lab test represents a column. For each entry in the table, we only chose the latest available test results after 2 months prior to the discovered date, as shown in figure 3. We adopted this selection since the data contained missing values. Thus, we mark entries having available test results as 'known', and 'unknown' otherwise. In addition, for the control class, the training set contains subjects whose earliest visit date is prior to January 1, 2014; while the test set contains subjects whose earliest visit date is after January 1, 2014. We also in the feature table for this class similar to the positive class.

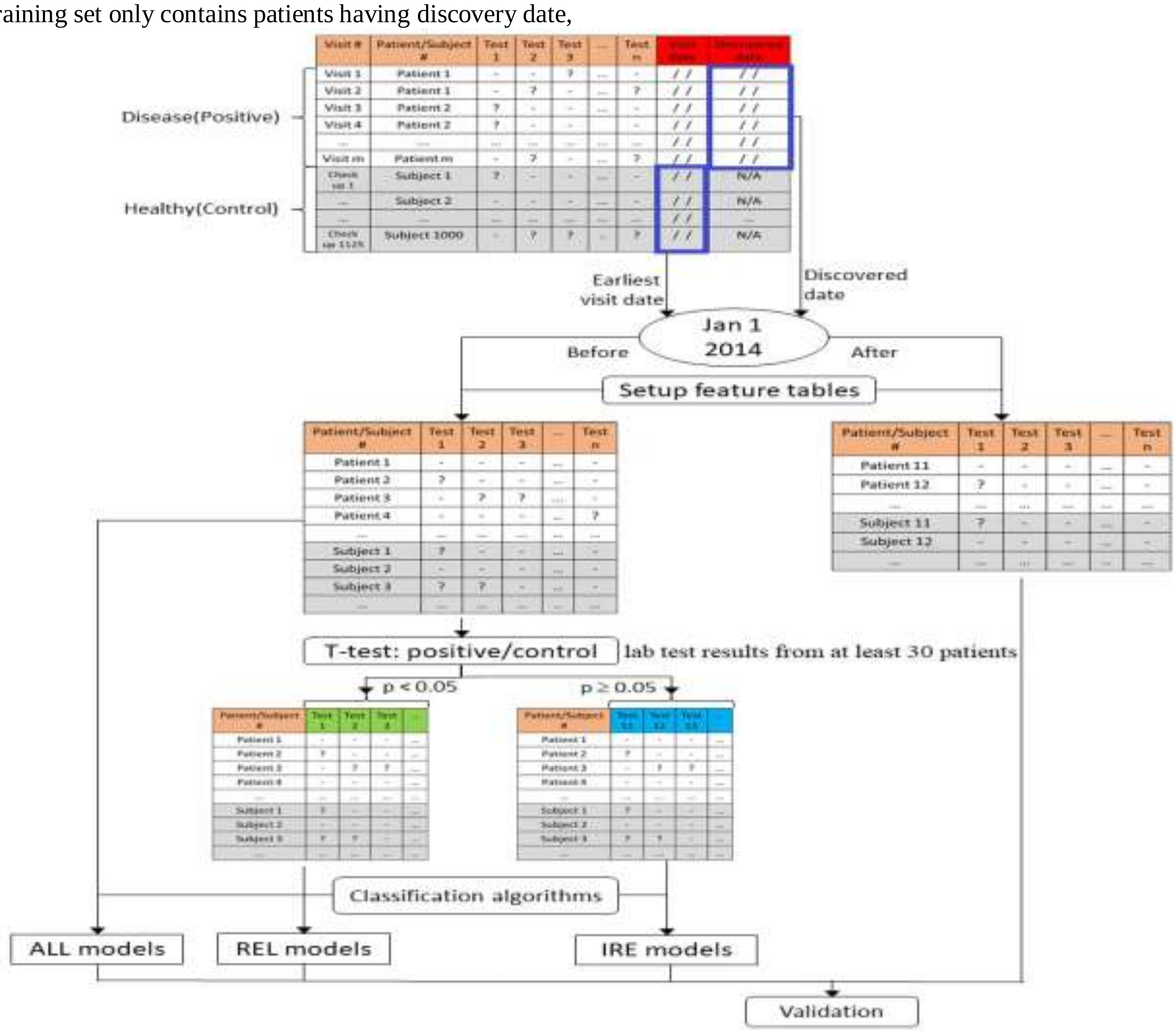

Fig 2 The overall framework in this paper: dividing the dataset into training and test set, setting up features table, finding association between disease and lab tests and validating the result by different classification models. Here, table entry '-' implies that the entry value is known; table entry '?' implies that the entry value is unknown.

We applied statistical and machine learning techniques to detect and validate the occurring diseases-lab test results associations. To mine the occurring diseases-lab test results associations, we apply the student t-test [18]. Since the data contains missing values, for each disease, we only compute a specific occurring diseases-lab test results association when there were at least 30 patients having the test results. As shown in figure 2, for each disease, we define that tests resulting in t-test p-value < 0.05 between the disease and the control classes are associated with the disease. More importantly, we only conducted the t-test using the training set. To validate these associations, we compared the disease-versus-control classification performance using three types of model. For the first type of model, noted as REL (abbreviation of relevant), we only use the disease's associated tests as features for classification. For the third type of model, noted as IRE (abbreviation of irrelevant), we only used the non-associated tests as features for classification. For the second type of model, noted as ALL, we used all tests to build the models. We trained the classification models using the training set and measure the performance on the test set, as shown in the above section. For training classification models, we applied Random Forest [15] implemented in Weka version 3.8 [19], which was significantly successful in Google's and Mt. Sinai's DeepPatient [10].

Disease X

Visit #	Patient #	Test 1	Test 2	Test 3	Visit date	Discovered date
Visit 1	Patient 1	0.1	1	10	Feb 5 2012	Jan 1 2013
Visit 2	Patient 1	0.2	2	?	Dec 24 2012	Jan 1 2013
Visit 3	Patient 1	0.3	3	?	Feb 25 2013	Jan 1 2013
Visit 4	Patient 1	0.4	?	?	Mar 1 2014	Jan 1 2013

Patient #	Test 1	Test 2	Test 3
Patient 1	0.4	3	?

Fig.3. A toy example of setting up the feature table with Disease X and Patient 1 having 4 visits. Here, Visit 1 is not used because the visit date is more than 2 months before X discovered date. Entries for column Test 1, Test 2 and Test 3 are the latest available test results after 2 months of X discovered date. '?' implies that the test result is unknown.

We also applied Support Vector Machine (SVM) [20], another popular machine learning technique for comparative purpose. Before executing the SVM classification, for each disease, we normalized the feature table as follow. For each test, from the 'known' entries, we used the z-score normalization to transform them. By the z-score normalization, the expected normalized test result is 0 [21]. This allows representing the 'unknown' entries in the feature table as 0 in Support Vector Machine.

Here, we organized the data into .arff files (compatible with Weka version 3.8 [18]. Each .arff file corresponds to one disease. Each disease has 6 .arff file: the training/test file for ALL, REL and IRE models.

III. RESULTS

A. *Significant associations between occurrence of chronic disease and results of lab tests*

We found 713 occurring diseases-lab test results associations with full details in Supplemental Table2. In figure 4, we demonstrate these association patterns in a heat map. Here, we sort the test (row) and disease (column) by the number of associations occurring in each test and disease. From the disease perspective, the fact that hyperlipidemia, diabetes, and hypertension stand among the top 3 diseases with the highest number of associations is not surprising, given that these diseases are among the greatest concern in China due to rapidly better living condition but poor lifestyle [22]. Interestingly, the red-blood-cell-related tests, including hemoglobin concentration and red blood cell count, show strong associations with most of the chronic diseases. The albumin-related tests show similar patterns to the red-blood-cell-related tests among the top 5 tests with the highest number of associations. Metabolic-related tests such as cholesterols, triglyceride, and glucose do not rank among the top 5 tests having the highest number of associations. This fact suggests that the general public in China should be more informed about the impact of red blood cell and albumin abnormality, which has been somewhat neglected in China due to the recent concerns on metabolic diseases (such as diabetes and hyperlipidemia) and lung cancer [23].

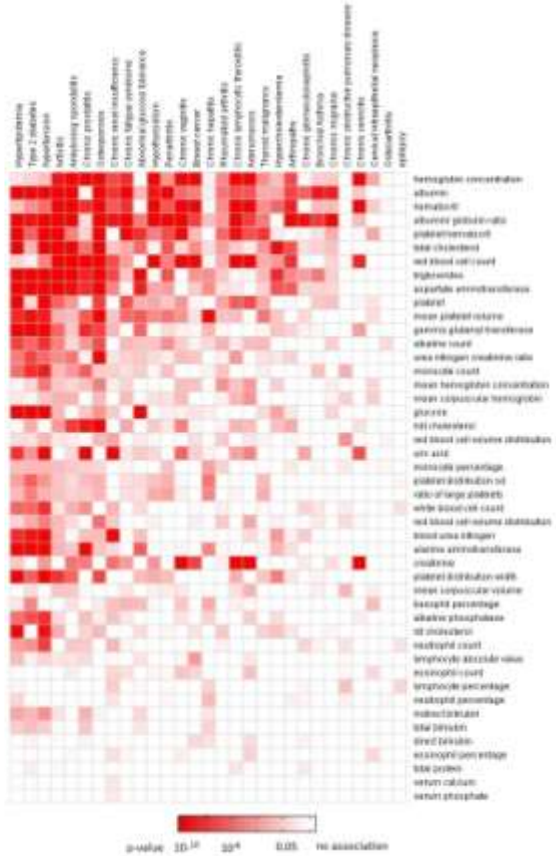

Fig.4. Heatmap illustrating the patterns of disease-tests association by p-value.

B. *Use associated tests as features improves the classification of disease*

Figure 5 shows that the classification models built upon only disease-associated tests using the Random Forest method (REL models) are completely superior to the models built upon only non-associated tests (IRE models). By average, the REL models achieve area-under-curve (AUC) of 0.931 and accuracy of 0.888; meanwhile, the IRE models only achieve AUC of 0.863 and accuracy of 0.831. The details of the classification result for each disease ICD could be found in supplemental Table 3. The REL models perform closely to the ALL models, where we use all test for disease classification (AUC: 0.953, Accuracy: 0.901). These facts validate the occurring diseases-lab test results associations found in the previous section. Classification using Support Vector Machine also shows that the REL models are superior to the IRE models. However, classification performance using linear Support Vector Machine is poorer. Supplemental table 4 is SVM's result. In figure 6, we showed that the models built from the top 15 tests, ordered by the p-value, would be sufficient to result in a good performance (AUC > 0.90).

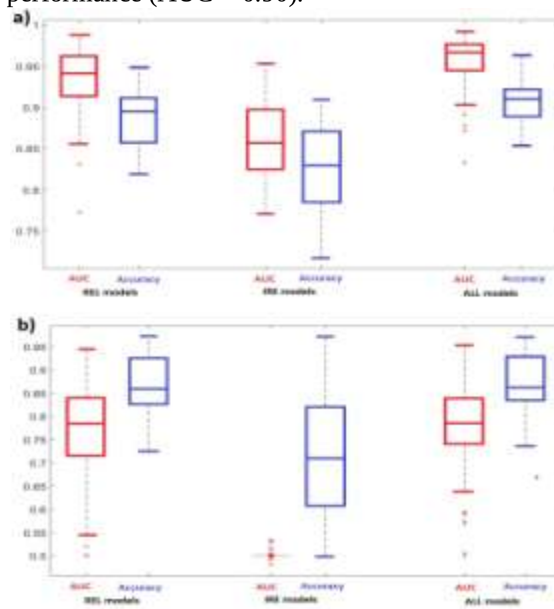

Fig.5. Comparison of classification performance among REL, IRE and ALL models built by a) Random Forest; b) Support Vector Machine.

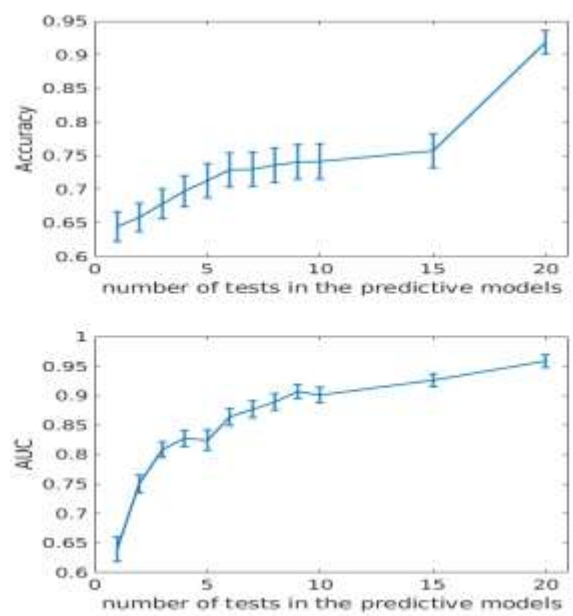

Fig.5. Prediction accuracy and AUC increased as the number of tested in the predictive model increases

We also compare our classification performance with the results from DeepPatient [10], which is among the latest state-of-the-art work in disease prediction using EHR data and using Random Forest. For all of the diseases overlapping between Deep Patient and our analysis, our REL models always show better performance. The details could be found in table 1.

TABLE 1 Classification results using Random Forest, in comparison with Deep Patient results.

Disease name	IRE models		ALL models		REL models		Deep Patient
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Breast cancer	0.878	0.878	0.963	0.920	0.938	0.903	0.762
Thyroid malignancy	0.816	0.763	0.948	0.889	0.932	0.847	N/A
Hypothyroidism	0.885	0.832	0.961	0.896	0.928	0.864	N/A
Chronic lymphocytic thyroiditis	0.896	0.812	0.938	0.858	0.903	0.819	N/A
Type 2 diabetes	0.883	0.778	0.991	0.933	0.988	0.925	0.907
Hypercholesterolemia	0.891	0.826	0.992	0.963	0.984	0.948	N/A
Hyperlipidemia	0.770	0.856	0.985	0.925	0.983	0.926	N/A
Chronic fatigue syndrome	0.877	0.809	0.976	0.914	0.961	0.891	N/A
Epilepsy	0.904	0.876	0.878	0.866	0.772	0.842	N/A
Chromic migraine	0.932	0.869	0.962	0.896	0.937	0.858	N/A
hypertension	0.857	0.908	0.987	0.938	0.980	0.933	0.574
Chronic obstructive pulmonary disease	0.784	0.816	0.872	0.879	0.867	0.870	0.688
Bronchial Asthma	0.953	0.875	0.967	0.918	0.903	0.858	N/A
Chronic hepatitis	0.894	0.828	0.943	0.879	0.917	0.837	N/A
Rheumatoid arthritis	0.855	0.786	0.974	0.914	0.957	0.897	N/A
Arthritis	0.857	0.774	0.976	0.914	0.960	0.895	N/A
Osteoarthritis	0.827	0.881	0.903	0.861	0.856	0.851	0.723
Arthropathy	0.838	0.737	0.973	0.921	0.962	0.904	N/A
Ankylosing spondylitis	0.926	0.860	0.972	0.904	0.942	0.881	N/A
Periarthritis	0.901	0.834	0.975	0.910	0.958	0.915	N/A
Osteoporosis	0.807	0.829	0.979	0.911	0.977	0.910	0.626
Chronic glomerulonephritis	0.947	0.910	0.973	0.955	0.964	0.946	N/A
Chronic renal insufficiency	0.789	0.867	0.832	0.907	0.831	0.903	N/A
Chronic prostatitis	0.774	0.717	0.946	0.854	0.934	0.847	N/A
Chronic cervicitis	0.951	0.908	0.974	0.903	0.918	0.908	N/A
Chronic vaginitis	0.838	0.780	0.959	0.886	0.956	0.894	N/A
Adenomyosis	0.832	0.829	0.960	0.929	0.949	0.895	N/A
Cervical intraepithelial neoplasia	0.811	0.867	0.891	0.898	0.890	0.867	N/A
Abnormal glucose tolerance	0.856	0.777	0.982	0.933	0.972	0.911	N/A

DISCUSSION

The high accuracy and AUC obtained in this paper shows that the health checkup could tell precisely whether a patient would have a specific chronic disease without further diagnostic procedures. With AI assistance, the checkup could potentially be as powerful as the diagnostic procedures in some diseases. For example, in diabetes diagnosis, Hirsch et al comprehensively examine diabetes definitions and estimate that using the test results directly from diabetes distribution would yield the classification AUC between 0.975 and 1 [24]. In this work, the checkup achieves AUC of 0.991 (ALL model) and 0.988 (REL model) in diabetes classification.

In this work, we have shown that mining occurring diseases-lab test results associations could not only improve the disease classification but also provide new insight on understanding disease risks. For example, the strong association between chronic metabolic-related diseases such as diabetes and hyperlipidemia and non-metabolic-related tests such as red blood cell and albumin could lead to new hypotheses for future studies. More importantly, table 1, which summarizes all occurring diseases-lab test results associations, includes only 46 tests, which is much less than the size of the core checkup package. Therefore, this result shows that we could reduce the size of the core checkup package such that the new core package preserves almost predictive capacity but costs significantly less. In the other hands, the classification performance using associated tests is high and better than the results showed in some state-of-the-art work. However, it is not necessary that the method presented in this work is better, because the work in [10] is completed in a more comprehensive data with longer duration, which allows leveling up the problem to predicting future disease occurrence.

The result of this work may provide some experience in handling the missing data in EHR. It is well-known that missing data is a critical issue in EHR analysis [25]. Thus, approximating the missing value has been considered the most important preprocessing step prior to disease classification and prediction. The Deep Patient [10] work is a typical example of handling missing value, in which the costly deep learning is applied only to estimate the missing values. However, this work shows that in disease classification, estimating missing values may not be critical, at least compared to the right combination of selecting which tests and which technique in the prediction. Here, our models built by Random Forest, for which we do not handle the missing value, show better performance than the model built by Support Vector Machine, in which the missing value issue is addressed.

We are aware of several limitations in this work. First, due to the data provider's and project requirement, we only have the data spanning within 4 years. The short duration and the limited number of caretaker (only 1AH) do not allow truly solving the future disease prediction problem. With longer data spanning time and more participating caretaker, we would be able to analyze follow-up checkups of subjects and bring the analytical techniques closer to a real-world application. Second, the data set is originally in Chinese; in addition, the data provider does not apply international standards to identify disease and lab tests fully. Therefore, translation and cleaning up the disease and lab test terminology must be done manually, which may be error-prone. Third, the data set does not contain the high number of chronic diseases and the patient coverage is not high. Therefore, the scope of our finding is limited to 713 occurring diseases-lab test results associations. Forth, due to the lack of integration with other types of health data and knowledge sources, we are not able to further annotate and categorize the occurring diseases and lab test results associations in this work. Abnormal results of some test are associated with occurrence of some diseases, of which some of them has known pathophysiological reason (like diabetes - glucose level). Some other associations may not be well-known. In these cases, it is need for further (literature) studies in all cases of the unexplained "associations" as without a proper literature check the empiric study results might be indistinguishable from arbitrary results, and unable to answer whether the occurring diseases-lab test results associations imply that the test is a new risk factor for the disease or the test and the disease are just co-occurring due to other factors. In addition, although the results may recommend may suggest a smaller 'core' checkup test from the original 97 tests, the ones being removed could be useful in detecting subtypes and severe conditions. A potential solution is creating more 'extended' packages from the tests being removed. The results and methodologies showed in this paper may just serve as the initial work for a future larger and more comprehensive study of the same topic at 1AH and the city of Wenzhou, Zhejiang, China, when we integrate them into a real-world Health IT application and collect the feedback from real patient-users.

IV. CONCLUSIONS

We do not intend to conclude the debate on the clinical benefits of health checkup, it shows another direction on how to use the checkup: by combining with health IT and AI/ML, the health checkup could be very powerful for predicting future chronic diseases, help to prevent these problems. In addition, an immediate conclusion is that at 1AH, we could design a new, smaller and cheaper 'core' checkup package while preserving the predictive power.

ACKNOWLEDGEMENT

This work is supported by Wenzhou Department of Science and Technology Development (Wenzhou Municipal Science and Technology Bureau), under project number ZG2017020, titled "Research and Development of Disease Prevention and Prediction System Based on Cloud Computing and Medical Big Data".

The authors especially thank Dr. Thanh Minh Nguyen from the Informatics Institute, the University of Alabama at Birmingham, for helpful comments in designing the study.

The authors thank IT staff member from the Department of Computer Technology and Information Management, The First Affiliated Hospital of Wenzhou Medical University, Zhejiang, China for assisting the authors in acquiring and preprocessing the data.

REFERENCES

[1] Krogsboll, L. T., K. J. Jorgensen, C. Gronhoj Larsen and P. C. Gotzsche (2012). "General health checks in adults for reducing morbidity and mortality from disease." Cochrane Database Syst Rev10: CD009009.

[2] Si, S., J. R. Moss, T. R. Sullivan, S. S. Newton and N. P. Stocks (2014). "Effectiveness of general practice-based health checks: a systematic review and meta-analysis." Br J Gen Pract64(618): e47-53.

[3] Suh, Y., C. J. Lee, D. K. Cho, Y. H. Cho, D. H. Shin, C. M. Ahn, J. S. Kim, B. K. Kim, Y. G. Ko, D. Choi, Y. Jang and M. K. Hong (2017). "Impact of National Health Checkup Service on Hard Atherosclerotic Cardiovascular Disease Events and All-Cause Mortality in the General Population." Am J Cardiol120(10): 1804-1812.

[4] Kudo, Y., T. Satoh, S. Kido, M. Ishibashi, E. Miyajima, M. Watanabe, T. Miki, M. Tsunoda and Y. Aizawa (2008). "The degree of workers' use of annual health checkup results among Japanese workers." Ind Health46(3): 223-232.

[5] Cao, X., J. Zhou, H. Yuan and Z. Chen (2015). "Cumulative effect of reproductive factors on ideal cardiovascular health in postmenopausal women: a cross-sectional study in central south China." BMC Cardiovasc Disord15: 176.

[6] Gu, D., P. Xu, Y. Yuan and H. Fu (2016). "Albuminuria is Suggested as a Potential Health Screening Biomarker for Senior Citizens and General Population with Hypertension or Diabetes in China." Clin Lab62(11): 2267-2269.

[7] Altman, D. G. (1990). Practical statistics for medical research, CRC press.

[8] Vapnik, V., S. E. Golowich and A. J. Smola (1997). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems.

[9] Kamkar, I., S. K. Gupta, D. Phung and S. Venkatesh (2015). "Stable feature selection for clinical prediction: exploiting ICD tree structure using Tree-Lasso." J Biomed Inform53: 277-290.

[10] Miotto, R., L. Li, B. A. Kidd and J. T. Dudley (2016). "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records." Sci Rep6: 26094.

[11] Van Der Maaten, L. and G. Hinton (2017). "Visualizing data using t-sne (2008)." J Mach Learn Res1117(9): 2579-2605.

[12] Weiskopf, N. G., G. Hripcsak, S. Swaminathan and C. Weng (2013). "Defining and measuring completeness of electronic health records for secondary use." J Biomed Inform46(5): 830-836.

[13] Weiskopf, N. G. and C. Weng (2013). "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research." J Am Med Inform Assoc20(1): 144-151.

[14] Dahlem, D., D. Maniloff and C. Ratti (2015). "Predictability Bounds of Electronic Health Records." Sci Rep5: 11865.

[15] Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." R news2(3): 18-22.

[16] "Bioscience research thriving in Wenzhou's Ouhai Life and Health Town."(2016) Retrieved 09/09/2017, from http://subsites.chinadaily.com.cn/ezhejiang/2016-09/28/c_58228.htm.

[17] ICD-10 online versions, World Health Organization.(2014)

[18] Peck, R., C. Olsen and J. L. Devore (2015). Introduction to statistics and data analysis, Cengage Learning.

[19] Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009). "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter11(1): 10-18.

[20] Vapnik, V., S. E. Golowich and A. J. Smola (1997). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems.

[21] Cheadle, C., M. P. Vawter, W. J. Freed and K. G. Becker (2003). "Analysis of microarray data using Z score transformation." J Mol Diagn5(2): 73-81.

[22] Martines, J. (2016). Rich Man, Poor Health: Class and Health in Modern China. The Diplomat.

[23] Hong, Q. Y., G. M. Wu, G. S. Qian, C. P. Hu, J. Y. Zhou, L. A. Chen, W. M. Li, S. Y. Li, K. Wang, Q. Wang, X. J. Zhang, J. Li, X. Gong, C. X. Bai, S. Lung Cancer Group of Chinese Thoracic and C. Chinese Alliance Against Lung (2015). "Prevention and management of lung cancer in China." Cancer121 Suppl 17: 3080-3088.

[24] Hirsch, A. G. and A. Scheck McAlearney (2013). "Measuring Diabetes Care Performance Using Electronic Health Record Data: The Impact of Diabetes Definitions on Performance Measure Outcomes." Am J Med Qual29(4): 292-299.

[25] Raghunathan, T. E. (2004). "What do we do with missing data? Some options for analysis of incomplete data." Annu Rev Public Health25: 99-117.

Author Details

A. Shuangquan Li, first author, majors in Biomedical Engineering, medical informatics.

B. Tongbin Zhang, second author, majors in Biomedical Engineering, medical informatics.

C. Chuandi Pan, *corresponding author. His main research direction is medical informatics.

D. Li Cai, fourth author. His main work is the application of information technology.