International Journal of Medical Informatics 97 (2017) 120–127

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

A machine learning-based framework to identify type 2 diabetes through electronic health records

Tao Zheng[a,b], Wei Xie[c], Liling Xu[b], Xiaoying He[d], Ya Zhang[a], Mingrong You[e], Gong Yang[e], You Chen (Ph.D)[f,*]

[a] Institute of Image Communication and Networking, Shanghai Jiao Tong University, Shanghai, China
[b] Tongren Hospital Shanghai Jiao Tong University, Shanghai, China
[c] Department of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, USA
[d] Department of Endocrinology, the First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China
[e] Division of Epidemiology, Vanderbilt University, Nashville, TN, USA
[f] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form
27 September 2016
Accepted 30 September 2016

Keywords:
Electronic health records
Type 2 diabetes
Data mining
Feature engineering
Machine learning

ABSTRACT

Objective: To discover diverse genotype-phenotype associations affiliated with Type 2 Diabetes Mellitus (T2DM) via genome-wide association study (GWAS) and phenome-wide association study (PheWAS), more cases (T2DM subjects) and controls (subjects without T2DM) are required to be identified (e.g., via Electronic Health Records (EHR)). However, existing expert based identification algorithms often suffer in a low recall rate and could miss a large number of valuable samples under conservative filtering standards. The goal of this work is to develop a semi-automated framework based on machine learning as a pilot study to liberalize filtering criteria to improve recall rate with a keeping of low false positive rate.
Materials and methods: We propose a data informed framework for identifying subjects with and without T2DM from EHR via feature engineering and machine learning. We evaluate and contrast the identification performance of widely-used machine learning models within our framework, including k-Nearest-Neighbors, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. Our framework was conducted on 300 patient samples (161 cases, 60 controls and 79 unconfirmed subjects), randomly selected from 23,281 diabetes related cohort retrieved from a regional distributed EHR repository ranging from 2012 to 2014.
Results: We apply top-performing machine learning algorithms on the engineered features. We benchmark and contrast the accuracy, precision, AUC, sensitivity and specificity of classification models against the state-of-the-art expert algorithm for identification of T2DM subjects. Our results indicate that the framework achieved high identification performances (~0.98 in average AUC), which are much higher than the state-of-the-art algorithm (0.71 in AUC).
Discussion: Expert algorithm-based identification of T2DM subjects from EHR is often hampered by the high missing rates due to their conservative selection criteria. Our framework leverages machine learning and feature engineering to loosen such selection criteria to achieve a high identification rate of cases and controls.
Conclusions: Our proposed framework demonstrates a more accurate and efficient approach for identifying subjects with and without T2DM from EHR.

1. Background and Significance

Type 2 diabetes mellitus (T2DM) is a major disease with high penetrance in humans around the globe, a trend that is still on the rise [1,2]. T2DM is a leading cause of morbidity and mortality and contributes to increased risks of heart disease by 2 to 4 times [1]. A significant number of research investigations have been devoted to it, notably by means of genome-wide association

* Correspondence to: 2525 West End Ave, Suite 1475, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37203 USA.
E-mail address: you.chen@vanderbilt.edu (Y. Chen).

http://dx.doi.org/10.1016/j.ijmedinf.2016.09.014
1386-5056/© 2016 Elsevier Ireland Ltd. All rights reserved.

study (GWAS) and phenome-wide association study (PheWAS) in hope of detecting more associations between genotypes and phenotypes [3–10,23–26,36]. To discover diverse genotype-phenotype associations affiliated with T2DM via PheWAS and GWAS, more cases (subjects with T2DM) and controls (subjects without T2DM) are required to be identified from electronic health records (EHR) [11,12,34,35].

A widely adopted approach for identifying subjects with and without T2DM is to have human experts (e.g., experienced physicians) manually design algorithms based on their experience and examination of EHR data [11,13–15]. However, such strategies increasingly prove to be limited and not scalable [11,13,15] due to the laborious process of human intervention and rule abstraction capabilities of experts. Furthermore, expert algorithms are often designed with conservative identification strategy, thus may fail to identify complex (e.g., borderline) subjects and miss a significant number of potential T2DM cases and controls. In research settings such as GWAS and PheWAS, accumulating large sample sizes is often highly desirable and discarding valuable samples will influence the potentiality to discover diverse genotype-phenotype associations [26,36]. A disease may be caused by the joint effects of multiple single nucleotide polymorphism (SNPs) (i.e. heterogeneity), while a SNP may lead to multiple diseases (i.e. pleiotropy) [32–34]. Involving more cases with diverse phenotypic characteristics such as comorbidities will enrich the association studies between phenotypes and genotypes. Given the limitations in high missing rate and laborious manual intervention, it is increasingly challenging for expert algorithms to scale to the ever-increasing volumes of diabetes related EHR data, secondary use and evolved GWAS and PheWAS studies [13,15,35].

Machine learning and data mining models are increasingly utilized in diabetes related research from EHR data (e.g., diabetes-related adverse drug effect, and association between periodontitis and T2DM) [27–29]. These studies have primarily focused on mining T2DM-related EHR data for clinical purposes, for instance, one such study aimed at forecasting clinical risk of diabetes from EHR [29]. The motivation and intended usage of the aforementioned work is different from ours, which aims to identify more cases and controls. Furthermore, the aforementioned study still has similar limitations in high missing rate [29]. To the best of our knowledge, very few studies have focused on reducing missing rate to identify more cases and controls for phenotyping purposes.

The goal of this work is to develop a semi-automated framework based on machine learning as a pilot study to identifying subjects with and without T2DM. Our method features two advancements: 1) low false positive rate; 2) high recall (i.e., detecting as many samples of interest as possible). To achieve these goals, we carefully approach feature engineering (i.e., construction of features for predictive modeling) by constructing representative features at three levels. We then train multiple popular machine learning models based on constructed features to identify cases and controls.

Our empirical evaluation is based on three years (ranging from 2012 to 2014) of EHR data from a large distributed EHR network consisting of multiple Chinese medical centers and hospitals in Shanghai, China. Our choice of this EHR repository is motivated by the fact that Chinese EHR data are often much worse than western EHR in terms of meaningful uses and data quality [18]. In addition, medical care in China often have non-standard unique procedures (such as wide adoption of traditional Chinese medicine) that are not represented in EHR and expert algorithms from elsewhere (such as from mainstream western counterparts), rendering standard or western expert algorithms less relevant. Given all such factors, the Chinese EHR repository provides an ideal test-bed for evaluating the accuracy and robustness of our proposed framework. In addition, the customization and empirical evaluation of a machine learning-based T2DM identification framework specifically for Chinese EHR is also of separate interest, which is under-explored despite constituting huge demand.

2. Research design and methods

2.1. Study materials

Our investigations in this work focus on three years of EHR data (from Year 2012 to 2014). The data was stored in our centered repository, which has been managed by the District Bureau of Health in Changning, Shanghai since 2008. The EHR data generated from 10 local EHR systems are automatically deposited into the centralized repository hourly.

We have 123,241 patients in total within the investigated three years. We use a filtering strategy to pre-select patients as our candidate samples whose EHR data are related to diabetes. We pre-selected samples whose EHRs should satisfy at least one of the three criteria: i) diabetes related diagnosis, ii) diabetes related medication and iii) diabetic laboratory test. Through this process, we managed to obtain 23,281 patient samples with diabetes related information. Our data preparation workflow is summarized in Fig. 1.

Our framework is based on supervised learning (e.g., classification, to be specific), which requires labeled training samples. Thus, we invited two clinical experts experienced in diabetes to assess EHRs of samples and label these samples into three categories: case, control and unconfirmed. We point out that, as is common for similar efforts, our expert review process is based on manually judging the whole record of each patient instead of only considering the selected few criteria in our data filtering or baseline expert algorithms (introduced later). Due to huge amount of manual effort in the expert reviewing process, as a pilot study, we randomly selected 300 samples out of the 23,281 pre-selected ones and concentrated our reviewing efforts on the smaller subset. For the investigated 300 selected samples, there are 20,384 records (e.g., diagnostic notes, communication notes and summary notes). Samples with two confirmed labels of T2DM from both clinicians will be considered as cases, samples with two confirmations of Non-T2DM considered as controls. The other samples with conflicting labels or two confirmations of un-determined from two clinicians will be denoted as unconfirmed ones. Through clinicians' assessments, we obtained 161 cases, 60 controls and 79 unconfirmed samples. For double check, we noticed that of the unconfirmed 79 samples, most (78.3%) are severely incomplete in their EHR documentation, which are not be suitable for EHR-based phenotyping. In order to reduce negative influences of incomplete EHRs on performances of our classification models, we dropped 79 unconfirmed samples.

Through our assessing processes of cases and controls, the separation range between cases and controls in our study is narrower than that in traditional expert algorithms as shown in Fig. 2.

This is because, in our study, controls refer to samples satisfying at least one of the following three criteria: i) one time of abnormal lab tests (HbA1C ≥6.0% or fasting plasma glucose ≥126 mg/dl or 2-h plasma glucose ≥200 mg/dl or random plasma glucose ≥200 mg/dl), ii) one time of prescribed diabetic medicine, and iii) one time of diabetic diagnosis. However, these controls were excluded in expert algorithms [11,13,15,30]. However, the widely used expert algorithms selected controls whose EHR data should not include any of the three above mentioned diabetic related information. The selection criteria of expert algorithms will miss many controls. For instance, we investigated a number of control samples, who had high values of HbA1C (≥6.0%) recorded, but their fasting and post-meal blood sugars were normal. Another example is we found several controls whose records contained prescriptions of diabetic medications, but no diabetic diagnoses and laboratory

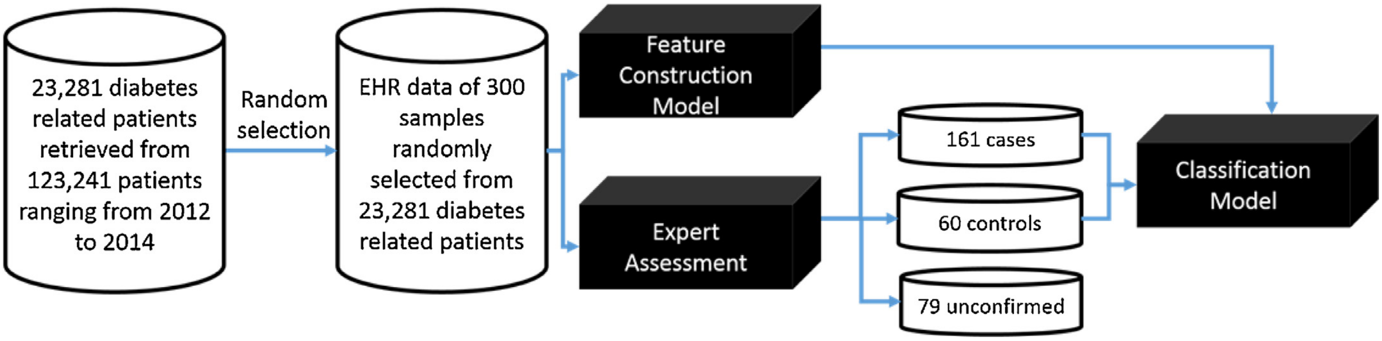
23,281 diabetes related patients retrieved from 123,241 patients ranging from 2012 to 2014
Random selection
EHR data of 300 samples randomly selected from 23,281 diabetes related patients
Feature Construction Model
Expert Assessment
161 cases
60 controls
79 unconfirmed
Classification Model

Fig. 1. A machine learning-based framework to identify subjects with and without T2DM from EHR data. The *case* refers to subjects with T2DM, and *control* refers to non-T2DM subjects.

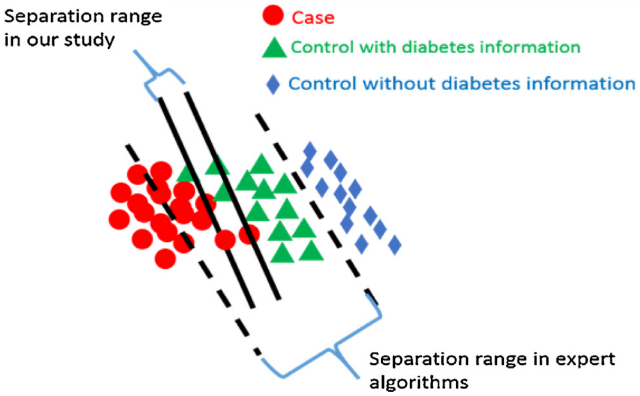
Separation range in our study
Case
Control with diabetes information
Control without diabetes information
Separation range in expert algorithms

Fig. 2. Separated boundary lines between cases and controls in our study and in traditional T2DM identification studies.

tests were found in their records. One of reasons is the medications they prescribed were not for themselves, but for their friends or someone else.

For the cases selection, expert algorithms selected samples whose EHR data should at least satisfy 2 of the following three requirements.

(1) Abnormal laboratory tests (glucose ≥110 mg/dl or HbA1c ≥6.0%)
(2) Diabetic medication
(3) Diabetic diagnosis

Such selection process does not consider patients satisfying no more than 1 of the above three requirements but considering as T2DM patients through their related support information such as diabetic complications and self-reported body weight loss, persistent hunger, polyuria, and polydipsia. As a result, these cases were missed. According to our selection criteria, the range of separations between cases and controls are much smaller than expert algorithms as shown in Fig. 2. We applied expert algorithms and our proposed framework to identify cases and controls in the same separation range (the range between two solid lines as shown in Fig. 2). Both types of algorithms were studied on the same sources of 300 samples.

Our proposed framework includes feature construction, and classification models. Feature construction transforms raw EHR data into statistical features, which are further served as input entities to feed classification models (as shown in Fig. 1). The expert algorithms extracted their three major features (abnormal laboratory tests, diabetic medication and diabetic diagnosis) in the same raw EHR data and then use their standards (workflows as depicted in the Fig. A1 of Appendix F) to identify cases and controls.

2.2. Feature construction

Constructing good features from EHR is often a must to warranty good prediction performance either for expert algorithms or machine learning-based models. This is because raw EHR data are often noisy, sparse, and contain unstructured information (e.g., text) that are not directly "computable". Traditional researches on identifying subjects with and without T2DM were using selection strategies built on three features: diabetic diagnosis, diabetic laboratory tests and diabetic medications extracted from EHRs of investigated samples [11,16]. Such researches are limited due to their high missing rates on identification of cases and controls. This is because such strategies applied a conservative selection criteria on cases and controls (e.g., satisfying two of the aforementioned three features) and were tested in a broader separation range between cases and controls (the range between two dashed lines as shown in Fig. 2).

In our work, we include borderline (samples between two dashed lines of Fig. 2), which can help to identify more cases and controls than traditional studies. To make the case/control identification more accurate, we need to incorporate more features than traditionally used. For instance, we constructed additional T2DM features such as self-reported diabetes related symptoms, and diabetic complications, and so on, in hope of better identifying borderline or more ambiguous samples. In total, we derived 110 features from seven sources (we denote this as First-Level features): "*demographic information*", "*communication report*", "*outpatient diagnosis report*", "*inpatient diagnosis report*", "*inpatient discharge summary*", "*prescription report*" and "*laboratory test report*", as summarized in Table 1 and with in-depth explanation off each feature in Table A1 of Appendix A.

Notably, the features includes supporting materials for T2DM such as diabetic complications (e.g., diabetic retinopathy, diabetic neuropathy, diabetic cerebral vascular and diabetic peripheral circulation diseases), self-reported symptoms (e.g., self-report of body weight loss, persistent hunger, polyuria, and polydipsia), additional Chinese traditional medications and more laboratory test items (e.g., two-hours, fast and random glucose tests).

For the features in the medication category, we list investigated medicines related with T2DM treatments as in Table A2 of Appendix B. Notably, to tailor for Chinese EHR, we added additional Chinese traditional medicine, and mixtures of Chinese traditional and western medicine into the medication list. This is due to observation that T2DM patients are usually treated with a combination of Chinese traditional and western medicine in China, which is different from

Table 1
First-level Features constructed from source "*demographic information*", "*communication reports*", "*outpatients diagnosis reports*", "*inpatients diagnosis reports*", "*inpatients discharge summaries*", "*prescription reports*" and "*laboratory test reports*".

Source	category	Feature
Demographic information		De-identification ID, age and gender of a subject. (feature ranging from f1 to f3 as shown in Table A1 of Appendix A)
Communication report	Self-reporting note	Number of times a subject reporting body weight loss, persistent hunger, polyuria, polydipsia, prescribed diabetes medicine or returning visits for diabetes in communication report. (feature ranging from f4 to f9 as shown in Table A1 of Appendix A)
	Diagnosis code	Number of times codes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation appeared in communication report. (feature ranging from f10 to f17 as shown in Table A1 of Appendix A)
	Diagnosis note	Number of times communication report containing notes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation disease. (feature ranging from f18 to f25 as shown in Table A1 of Appendix A)
Outpatient diagnosis record	Diagnosis code	Number of times codes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation were appeared in outpatient diagnosis record. (feature ranging from f26 to f33 as shown in Table A1 of Appendix A)
	Diagnosis note	Number of times outpatient diagnosis record containing notes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation disease. (feature ranging from f34 to f41 as shown in Table A1 of Appendix A)
Inpatient discharge summary	Diagnosis notes	Number of times inpatient discharge summary containing notes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation disease. (feature ranging from f42 to f49 as shown in Table A1 of Appendix A)
Inpatient diagnosis record	Diagnosis codes	Number of times codes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation were appeared in inpatient diagnosis record. (feature ranging from f50 to f57 as shown in Table A1 of Appendix A)
	Diagnosis notes	Number of times inpatient diagnosis record containing notes of type 2 diabetes, diabetic retinopathy, diabetic neuropathy, diabetic eye disease, diabetic kidney disease, diabetic cerebral vascular disease or diabetic peripheral circulation disease. (feature ranging from f58 to f65 as shown in Table A1 of Appendix A)
Prescription record	Medication	Number of prescriptions appearing in prescription report for oral hypoglycemic, insulin, Chinese traditional hypoglycemic, a mixture of western and Chinese traditional oral hypoglycemic, Epalrestat, Alpha-glucosidase inhibitor, Dipeptidyl peptidase IV(DPP-IV) inhibitors, Meglitinides, Sulfonylureas, Thiazolidinedione, Biguanides, Incretin Mimetics, GLP-1 (glucagon-like peptide 1) mimetics, compounds of sulfonylurea and thiazolidinedione, compounds of Biguanides and Dipeptidyl peptidase IV(DPP-IV) inhibitors, compounds of Biguanides and Sulfonylureas, or compounds of Biguanides and Thiazolidinedione. (feature ranging from f66 to f82 as shown in Table A1 of Appendix A)
Laboratory test report	Venous plasma glucose test	Number of times for 2-h venous plasma glucose test, 2-h peripheral plasma glucose test ≥11.1 mmol/l (200 mg/dl), fasting venous plasma glucose test, fasting venous plasma glucose test ranging from 6.1 to 7.0 mmol/l (110 and 126 mg/dl), random venous plasma glucose test, or random venous plasma glucose test ≥11.1 mmol/l (200 mg/dl). (feature f83, f84, f87, f88, f91, and f92 as shown in Table A1 of Appendix A) The maximum value of 2-h venous plasma glucose test, fasting venous plasma glucose test, or random venous plasma glucose test. (feature f85, f89, and f93 as shown in Table A1 of Appendix A) The minimum value of 2-h venous plasma glucose test, fasting venous plasma glucose test, or random venous plasma glucose test. (feature f86, f90, and f94 as shown in Table A1 of Appendix A)
	Peripheral plasma glucose test	Number of times for 2-h peripheral plasma glucose test, 2-h peripheral plasma glucose test ≥11.1 mmol/l (200 mg/dl), fasting peripheral plasma glucose test, fasting peripheral plasma glucose test ranging from 6.1 to 7.0 mmol/l (110 and 126 mg/dl), random peripheral plasma glucose test, or random peripheral plasma glucose test ≥11.1 mmol/l (200 mg/dl). (feature f95, f96, f99, f100, f103, and f104 as shown in Table A1 of Appendix A) The maximum value of 2-h peripheral plasma glucose test, fasting peripheral plasma glucose test, or random peripheral plasma glucose test. (feature f97, f101, and f105 as shown in Table A1 of Appendix A) The minimum value of 2-h peripheral plasma glucose test, fasting peripheral plasma glucose test, or random peripheral plasma glucose test. (feature f98, f102, and f106 as shown in Table A1 of Appendix A)
	HbA1C test	Number of times for HbA1c test, HbA1C test ≥ 6.5%. (feature f107, and f108 as shown in Table A1 of Appendix A) The maximum value of HbA1C test (feature f109 as shown in Table A1 of Appendix A) The minimum value of HbA1C test (feature f110 as shown in Table A1 of Appendix A)

the common practice (i.e., western medicine only) of the western world and was thus neglected by western EHR-oriented studies.

For the diagnosis notes related features, we use regular expressions combining positive notes and negative notes as depicted in Table A3 of Appendix C to build each of them.

2.3. Feature summarization

Features (as shown in Table A1 of Appendix A) cover seven EHR sources, however, some sources have the same type of features. For instance, f_{10} in the source of "*communication report*", f_{26} in "*outpatient diagnosis record*", and f_{50} in "*inpatient diagnosis record*" have the same definition on the counting of diagnosis codes. These features are highly correlated with each other, which will influence performances of computational models to do classification [17,19,22]. And thus we merge correlated features into one feature by summarizing them. For instance, f_{10}, f_{26} and f_{50} are summarized as a new feature $f'_{10} = f_{10} + f_{26} + f_{50}$, which represents the total number of times T2DM diagnosis codes appearing in "*communication reports*" (f_{10}), "*outpatient diagnose records*" (f_{26}) and "*inpatient diagnose records*" (f_{50}) respectively. By using the same way, we summarize all similar features across the seven sources into 36 features as shown in Table A4 of Appendix D. At the same time, the features within a source are also correlated, so we transform 36 features into final 8 features through summarizing correlated features within a source. The final 8 features are listed in Table A5 of Appendix E.

2.4. Classification

We use several widely-used classification model such as k-Nearest-Neighbors (kNN), Naïve Bayes (NB), Decision Tree (J48), Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) to model patterns of cases and controls based on our extracted features and then use the models to test the ability of our extracted features on identifications of T2DM subjects. These classification models are frequently utilized in a wide range of fields, and are recognized as popular choices for classification tasks [20,21,37].

3. Results

3.1. Experimental set-up

Our framework adopts feature engineering by abstracting the EHR data at three different levels. This ensures to leverage more available data while maximizing predictive power. For the 221 T2DM samples (160 cases and 61 controls) collected with expert labels, we first construct 110 features (as mentioned before; see also Table A1 in Appendix A) to represent their EHR data. This is roughly a summarized and structured version of the raw EHR data. To prevent data sparsity and noise, we then derive higher-level features by condensing the data into 36 features (see Table A4 in Appendix D) and 8 features (see Table A5 in Appendix E), respectively. Such abstraction is mainly based on common knowledge of EHR data hierarchies.

In our framework, we apply several widely-used machine learning models, including kNN, NB, J48, RF, SVM and LR. The goal is to find out the comparative performance of machine learning models against expert algorithms. We used Weka package to apply these models on our engineered features [31]. We perform training and evaluation on different abstraction levels of feature sets, e.g., on the 107 aforementioned features[1] (the first level of features; as shown in Table A1 of Appendix A), 33 features (the second level of features; as shown in Table A4 of Appendix D), and 5 features (the third level of features; as shown in Table A5 of Appendix E), respectively. We conduct extensive comparison of different classifiers on the same level of features, as well as performance across the three different levels of feature sets described before. Furthermore, we use the state-of-the-art expert algorithm [11] as a benchmark baseline, which is widely adopted by several large EHR and genetics consortia studies. We emphasize that the expert algorithm [11] is evaluated on the same raw EHR data as mentioned before.

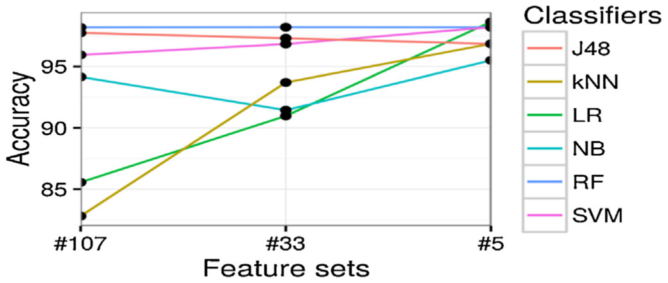

Fig. 3. Prediction accuracy (y-axis) with different feature sets (x-axis), categorized by different classifiers (different lines plotted).

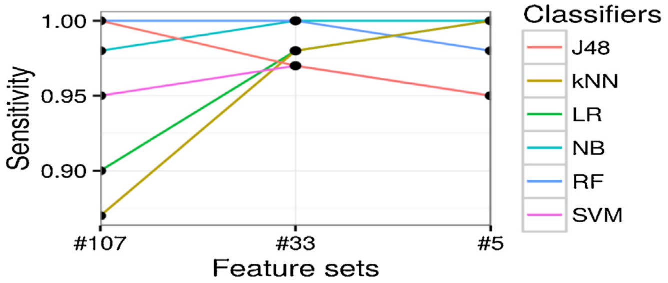

Fig. 4. Prediction sensitivity [True positive rate] (y-axis) with different feature sets (x-axis), categorized by different classifiers (different lines plotted).

We also point out that our primary focus of this work is to demonstrate feasibility/suitability of machine learning-based framework for the given task, and to provide general model recommendations and suggestions. Comprehensive and systematic benchmark of different machine learning models is not the main focus and is a separate topic with extensive literature. To keep our work focused and data-efficient, we adopt default recommended model parameters instead of performing hyperparameter tuning, since the latter often requires setting aside independent validation datasets, which may not be a wise option given our relatively small (and valuable) expert-labeled dataset. Our decision thresholds in certain models are also based on default configurations in Weka software [31]. For instance, in logistic regression, we use $p = 0.50$ as the classification cut-off.

3.2. Performance of classification models

For each classifier and each level of feature set, we conduct 4-fold cross-validation and report on the average performance and standard deviation. We demonstrate the prediction accuracy results in Fig. 3, which measures the ratio of correctly predicted samples. In Fig. 4, the prediction sensitivity (also called recall) results are reported, which measures the ratio of true positives against all positives. Lastly, in Fig. 5, we plot the specificity, which denotes the proportion of true negatives of all negatives. The precision (or positive predictive value) results are illustrated in Fig. 6. For more comprehensive comparison, we also present the area under the receiver operating characteristic (ROC) curve (AUC) in Fig. 7, which demonstrates the trade-off between false positive and true positive rates (larger AUC generally implies better performance). All detailed numerical metrics are also summarized in Table 2.

Based on the above results, J48, RF, and SVM have high prediction performances across various metrics, yielding over 0.95 in

[1] We exclude three demographic features, because they do not indicate any significant differences between cases and controls, and in contrast, they will influence the correct determinations of classifiers such as kNN.

Table 2
Comparison of different classifiers and the expert algorithm (baseline), measured by their average performance (and standard deviation) in cross-validation.

Classifiers	Feature Sets	Accuracy	Sensitivity	Specificity	Precision	AUC
Expert Algorithm	–	0.84	0.78	1.00	1.00	0.71
LR	**#107**	0.86 (0.06)	0.90 (0.09)	0.84 (0.10)	0.70 (0.11)	0.88 (0.07)
	#33	0.91 (0.04)	0.98 (0.03)	0.88 (0.06)	0.77 (0.07)	0.92 (0.03)
	#5	**0.99 (0.01)**	**1.00 (0)**	0.98 (0.01)	0.95 (0.03)	0.99 (0.01)
NB	**#107**	0.94 (0.05)	0.98 (0.03)	0.93 (0.07)	0.85 (0.11)	0.98 (0.02)
	#33	0.91 (0.07)	1.00 (0)	0.88 (0.10)	0.79 (0.15)	1.00 (0)
	#5	0.96 (0.03)	**1.00 (0)**	0.94 (0.05)	0.87 (0.09)	**1.00 (0)**
RF	**#107**	0.98 (0.01)	1.00 (0)	0.97 (0.02)	0.94 (0.05)	1.00 (0)
	#33	0.98 (0.01)	1.00 (0)	0.97 (0.02)	0.94 (0.05)	1.00 (0)
	#5	0.98 (0)	0.98 (0.03)	0.98 (0.01)	0.95 (0.03)	**1.00 (0)**
kNN	**#107**	0.83 (0.06)	0.87 (0.05)	0.81 (0.08)	0.65 (0.09)	0.91 (0.01)
	#33	0.94 (0.05)	0.98 (0.03)	0.92 (0.08)	0.84 (0.12)	0.98 (0.02)
	#5	0.97 (0.03)	**1.00 (0)**	0.96 (0.04)	0.90 (0.08)	0.99 (0.01)
SVM	**#107**	0.96 (0.04)	0.95 (0.03)	0.96 (0.04)	0.91 (0.10)	0.96 (0.03)
	#33	0.97 (0.02)	0.97 (0.04)	0.97 (0.02)	0.93 (0.06)	0.97 (0.02)
	#5	0.98 (0.01)	0.95 (0.03)	**0.99 (0.01)**	**0.98 (0.03)**	0.97 (0.02)
J48	**#107**	0.98 (0.02)	1.00 (0)	0.97 (0.02)	0.93 (0.05)	0.98 (0.01)
	#33	0.97 (0.02)	0.97 (0.04)	0.97 (0.02)	0.94 (0.05)	0.99 (0.01)
	#5	0.97 (0.03)	0.95 (0.03)	0.97 (0.03)	0.94 (0.07)	0.98 (0.03)

The bold values indicate the best models in terms of accuracy, sensitivity, specificity, precision and AUC. The significance values are inappropriate for them.

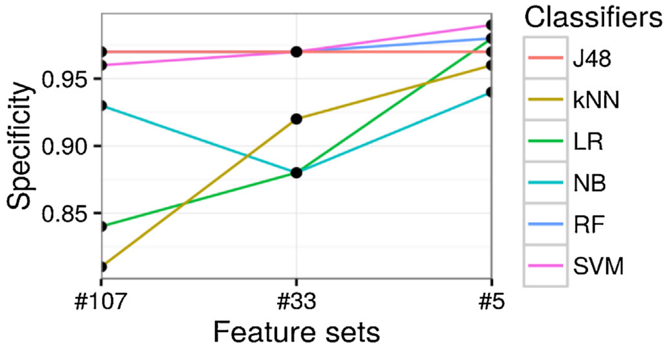

Fig. 5. Prediction specificity [True negative rate] (y-axis) with different feature sets (x-axis), categorized by different classifiers (different lines plotted).

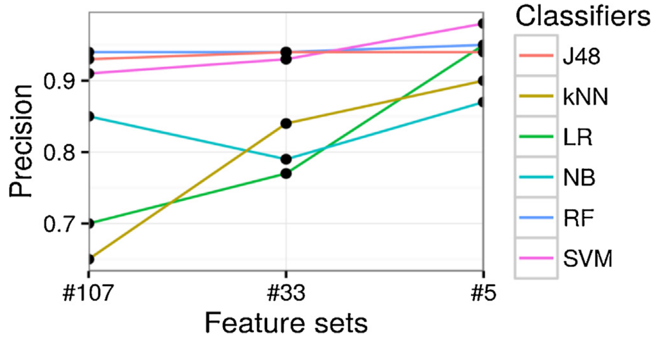

Fig. 6. Prediction precision [Positive predictive value] (y-axis) with different feature sets (x-axis), categorized by different classifiers (different lines plotted).

accuracy, sensitivity, specificity, and AUC on all three levels of features. As a comparison, the state-of-the-art expert algorithm [11] leads to performance of 0.84 in accuracy, 0.78 in sensitivity, 1.00 in specificity, and 0.71 in AUC. This indicates that our features constructed at all the three levels can identify T2DM subjects much better than the popular expert algorithm. State-of-the-art expert algorithm performs slightly better (and almost perfectly) in terms of specificity (1.00) and precision (1.00). This seems most likely due to its stringent conditions on case selection (e.g., a case subject should satisfy any two of the three metrics: diabetic diagnosis, diabetic medications and diabetic lab tests). Obviously, none of our controls satisfy the requirements of cases set above via expert algorithm, and thus bringing the specificity of the expert algorithm to 1 in our experiments.

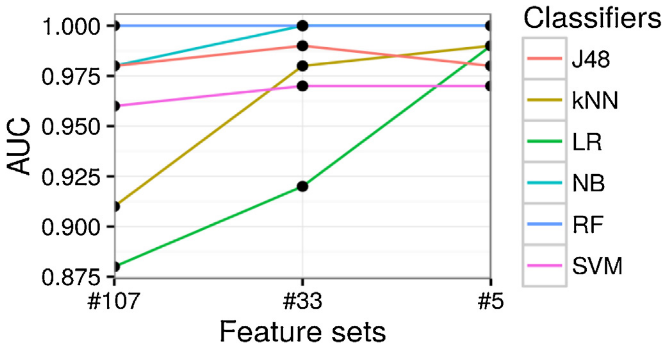

Fig. 7. Prediction AUC (y-axis) with different feature sets (x-axis), categorized by different classifiers (different lines plotted).

LR has the highest accuracy (0.99) at the third level of features (as shown in Fig. 2 and Table 2), with several other models closely following its performance, such as RF and SVM (with 0.98 in accuracy).

In terms of sensitivity, according to Fig. 4 and Table 2, most models experience performance improvement as features get summarized into higher levels. This indicates that simple feature engineering can boost the performance. Meanwhile, multiple models (e.g., LR, NB, and kNN) achieve (near) perfect sensitivity, namely 1.00, on the second or third levels of feature sets. This implies that our framework is highly efficient in make full use of all available data, especially regarding valuable case subjects which are often much rarer and more difficult to accumulate for subsequent studies (such as GWAS and PheWAS).

Accuracy and sensitivity of all classifiers at the third level of features are more stable than on the other two levels of features as shown in Figs. 3 and 4, which indicates the summarized final 5 features are stable discriminators to identify T2DM subjects.

For specificity shown in Fig. 5 and Table 2, half of the classifiers have performance greater than 0.95 (e.g., RF, SVM, and J48). LR and kNN performance worst when leveraging the first level of features.

This may be due to sparsity of features (thus many features end up being noise) or correlated features, which can bias such classifiers.

The AUC performance of our framework also exhibits similar encouraging results. In brief, when trained over second- or third-level feature sets, all classifiers (except LR) manage to perform well above 0.95, which is significantly better than random guessing (0.50) and almost approaching the perfect 1.00. State-of-the-art expert algorithm [11] only scored 0.71 in AUC, making it significantly worse than all models in our framework.

Roughly speaking, as is demonstrated across different metrics in Figs. 3–7, there is a general trend of increasing predictive performance, as features are abstracted into higher levels. This demonstrates the importance of our feature engineering approach. In addition, we observe better performance improvement from feature engineering than from choices of different machine learning models. This implies that when sample sizes are not sufficiently large (as in most EHR settings), a better strategy to maximize performance should be to refine features.

Overall, across all major metrics, models such as RF, J48 and SVM are more stable than the other three classifiers (kNN, NB, and LR) across the three levels of features. This may be because RF, J48 and SVM are less influenced by sparsity and noise of EHR data, whereas LR, kNN and NB are more vulnerable to these issues.

4. Discussion

Traditional expert algorithms use a wide range of separation to select cases and controls, and as a result, a large number of cases and controls are missed. In order to reduce missing rates of current studies, we propose a machine learning-based framework to identify cases and controls in a narrower separation range. We evaluated our framework through Chinese EHR data, and the experimental results show our framework can achieve higher performances than the state-of-the-art algorithm in such EHR data.

However, this work is a pilot study, which is limited in the following aspects:

Firstly, the number of samples (cases and controls) we studied needs to be enlarged in future. Although current selected 221 samples achieve high identification rates on detecting cases and controls, we still need more samples from our repository to confirm scalability of our models. For instance, we can use our classification models to select candidate cases and controls from 23,281 diabetes related patients, and then submit them to clinicians for reviewing. Under such semi-supervised way, we can gain more samples to enrich our framework via a large scale of training (e.g., on more diverse cases and controls) and testing (e.g., on independent new unseen samples). This process will require more reviewing efforts from humans, and will be considered as our next plan.

Second, our framework still involves human efforts in designing of features and confirmations of cases and controls. Although we spent a large amount of time on designing of features, we believe our extracted features could be utilized in other related studies without involving human efforts, which could save them huge amount of time. The evaluations of cases and controls are used to feed our machine learning models. According to achieved high performances of our classification models, researchers can use our model to select cases and controls with a high accuracy, which will save them time to get cases and controls through expert assessments.

Third, compared with expert algorithms in terms of high specificity (small number of non-T2DM are considered as T2DM), our models have lower specificity. This is because we include most of patients between the separation range of cases and controls in expert algorithms (the range between two dotted lines as shown in Fig. 2), and as a result, it is hard to make sure all selected cases are predicted correctly. If a study focuses on accuracy of T2DM patient identification more than on number of T2DM patients required, then expert algorithms would be a better choice. If number of cases and controls has higher priority, then our framework would be a better choice.

Fourth, our framework is not confirmed on EHR data of other institutes such as western EHR data. Although the framework achieves a high performance on Chinese EHR data, we believe such EHR data based strategy is also fit for identifying T2DM subjects on western EHR data, and we will test such hypothesis in our next step.

Finally, our methodology focuses on case/control design for traditional association study between phenotypes and genotypes, which requires a perfect precision (wide range of separation between cases and controls in expert algorithms as shown in Fig. 2). The reduced precision rate (leading to a higher recall rates of cases and controls) of our method may influence the traditional association studies. However, as the development of computational phenotyping from EHR data, the association studies will involve more cases with diverse phenotypic characteristics such as comorbidities to enrich the association studies between phenotypes and genotypes. This is because, a disease may be caused by the joint effects of multiple SNPs (i.e. heterogeneity), while a SNP may lead to multiple diseases (i.e. pleiotropy).

5. Conclusions

Identifying subjects with and without T2DM is the first step to enable subsequent analysis such as GWAS and PheWAS. In this work, we propose an accurate and efficient framework as a pilot study to identify subjects with and without T2DM from EHR data. Our framework leverages machine learning to automatically extract patterns of T2DM. And we further boost its predictive power by overcoming the wide separation rage of cases and controls in expert algorithms. Our feature engineering framework considers a diverse set of data features spanning diabetic diagnosis codes, diagnosis notes, complications, self-reports, medications (both standard and traditional Chinese medicine), and laboratory tests to represent diabetes related patients. Based on engineered features, we train classification models. We collected 160 T2DM cases and 61 controls and use 4-fold cross validation strategy to evaluate performances of classification models. The experimental results show that our framework can identify subjects with and without T2DM at an average AUC of around 0.98, significantly outperforming the state-of-the-art at an AUC of 0.71.

Author contributions

T. Zheng, W. Xie, and L. Xu designed the algorithm, analyzed the results of the experiments, and wrote the paper. X. He, Y.Zhang, M.You, and G.Yang analyzed the results of the experiments, and revised the paper. Y. Chen designed and supervised the study, analyzed the data and the results of the experiments, and wrote the paper.

Funding

This project is supported in part by the National High Technology Research and Development Program of China (2013AA020418) and NIH grantR00LM011933.

Competing interests statement

The authors have no competing interests to declare.

Summary points

What was already known on this topic?

(1) Electronic health record (EHR) data can be leveraged to distinguish patients with and without type 2 diabetes Mellitus (T2DM).

(2) (Existing EHR-based algorithms often suffer in terms of accuracy and could miss a large number of valuable patients under conservative filtering standards.

What this study added to our knowledge?

(1) Identification accuracy of T2DM patients could be improved by loosen conservative selecting criteria of cases (patients with T2DM) and controls (patients without T2DM) in current algorithms.

(2) A machine learning-based framework to Identify T2DM subjects.

Acknowledgements

We thank members from the Changning regional distributed EHR network, Shanghai, China, for sharing of de-identified EHR data studied in this paper. We would like to thank physicians from hospitals and medical centers belonging to Changning district to evaluate cases and controls.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2016.09.014.

References

[1] Centers for Disease Control and Prevention, National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, US Department of Health and Human Services, Atlanta, GA, 2014.

[2] Y xu, L, wang, J He prevalence and control of diabetes in chinese adults, JAMA 310 (9) (2013) 948–959.

[3] W. Yang, W. Zhao, J. Xiao, Medical care and payment for diabetes in China: enormous threat and great opportunity, PLoS One 7 (2016) e39513.

[4] p. Diabetes Prevention Program Research Grou, Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin, N. Engl. J. Med. 346 (6) (2002) 393–403.

[5] M.T. Frayling, Genome–wide association studies provide new insights into type 2 diabetes aetiology, Nat. Rev. Genet. 8 (2007) 657–662.

[6] E.S. Lander, et al., The common PPARbig gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes, Nat. Genet. 26 (2000) 76–80.

[7] UK Prospective Diabetes Study (UKPDS) Group, Effect of intensive blood-glucose control with metformin on complications in overweight subjects with type 2 diabetes (UKPDS 34), Lancet 352 (9131) (1998) 854–865.

[8] C.N. Hales, D.J.P. Barker, Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis, Diabetologia 35 (1992) 595–601.

[9] C.N. Hales, D.J. Barker, The thrifty phenotype hypothesis Type 2 diabetes, Br. Med. Bull. 60 (2001) 5–20.

[10] S. Bo, et al., Hyperuricemia and hyperuricemia in type 2 diabetes: two different phenotypes, Eur. J. Clin. Invest. 31 (2001) 318–321.

[11] A.N. Kho, et al., Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study, J. Am. Med. Inform. Assoc. 19 (2) (2012) 212–218.

[12] R.W. Grant, et al., Practice-linked online personal health records for type 2 diabetes mellitus, JAMA Intern. Med. 168 (16) (2008) 1776–1782.

[13] R. Kudyakov, J. Bowen, E. Ewen, S.L. West, Y. Daoud, N. Fleming, A. Masica, Electronic health record use to classify subjects with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management, Popul. Health Manage. 15 (1) (2012) 3–11.

[14] M.L. Ho, N. Lawrence, C. Walraven, The accuracy of using integrated electronic health care data to identify subjects with undiagnosed diabetes mellitus, J. Eval. Clin. Pract. 18 (3) (2012) 606–611.

[15] W.Q. Wei, et al., The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects, Int. J. Med. Inf. 82 (2013) 239–247.

[16] W.Q. Wei, et al., Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus, J. Am. Med. Inform. Assoc. 19 (2) (2012) 219–224.

[17] Y. Chen, Y. Li, X.Q. Cheng, L. Guo, Survey and taxonomy of feature selection algorithms in intrusion detection system, in: Conference on Information Security and Cryptology, LNCS4318, 2006, pp. 153–167.

[18] J. Lei, P. Sockolow, P. Guan, Q.J. Zhang, A comparison of electronic health records at two major Peking University Hospitals in China to United States meaningful use objectives, BMC Med. Inform. Decis. Mak. 13 (2013) 96.

[19] B.K. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (2) (1995) 227–234.

[20] S. Dreiseitl, L.O. Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (5–6) (2002) 352–359.

[21] J.L. Lustgarten, V. Gopalakrishnan, H. Grover, S. Visweswaran, Improving classification performance with discretization on biomedical datasets, AMIA Ann. Symp. Proc. 44 (2008) 5–449.

[22] J. Hua, Z. Xiong, J. Lowey, E. Suh, E.R. Dougherty, Optimal number of features as a function of sample size for various classification rules, Bioinformatics 21 (8) (2005) 1509–1515.

[23] F.S. Collins, H. Varmus, A new initiative on precision medicine, N. Engl. J. Med. 372 (2015) 793–795.

[24] B.F. Voight, et al., Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis, Nat. Genet. 42 (7) (2010) 579–589.

[25] L.K. Billings, J.C. Florez, The genetics of type 2 diabetes: what have we learned from GWAS? Ann. N. Y. Acad. Sci. 1212 (1) (2010) 59–77.

[26] J.C. Denny, et al., Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data, Nat. Biotechnol. 31 (12) (2013) 1102–1111.

[27] M. Marinov, A. Mosa, I. Yoo, S.A. Boren, Data-mining technologies for diabetes: a systematic review, J. Diabetes Sci. Technol. 5 (6) (2011) 1549–1556.

[28] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from EMR data using machine learning, AMIA Ann. Symp. Proc. 2012 (2012) 606–615.

[29] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients' data, Artif. Intell. Med. 41 (3) (2007) 251–262.

[30] C. Shivade, P. Raghavan, E. Fosler-Lussier, et al., A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. 21 (2) (2014) 221–230.

[31] C.T. Smith, E. Frank, Statistical Genomics: Methods and Protocols, Chapter Introducing Machine Learning Concepts with WEKA, Springer, New York, 2016, pp. 353–378.

[32] N. Solovieff, C. Cotsapas, P. Lee, S.M. Purcell, J.W. Smoller, Pleiotropy in complex traits: challenges and strategies, Nat. Rev. Genet. 14 (2013) 483–495.

[33] Q. Peng, J. Zhao, F. Xue, PCA-based bootstrap confidence interval tests for gene-disease association involving multiple SNPs, BMC Genet. 11 (2010) 6.

[34] Y. Wang, P. Sung, P. Lin, Y. Yu, R. Chung, A multi-SNP association test for complex diseases incorporating an optimal P-value threshold algorithm in nuclear families, BMC Genomics 16 (2015) 381.

[35] P. Mitteroecker, J.M. Cheverud, M. Pavlicev, Multivariate analysis of genotype-phenotype association, Genetics 203 (2016) 3.

[36] W. Xie, et al., SecureMA: protecting participant privacy in genetic association meta-analysis, Bioinformatics 30 (23) (2014) 3334–3341.

[37] W. Li, et al., Supporting regularized logistic regression privately and efficiently, PLoS One 11 (6) (2016) e0156479.