Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study

Ariana E. Anderson [a], Wesley T. Kerr [a,b,*], April Thames [a], Tong Li [a], Jiayang Xiao [a], Mark S. Cohen [a,c,d]

[a] Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, United States
[b] Department of Biomathematics, David Geffen School of Medicine at UCLA, United States
[c] Departments of Psychology, Neurology, Radiology, Biomedical Engineering, Biomedical Physics, University of California, Los Angeles, United States
[d] California NanoSystems Institute, University of California, Los Angeles, United States

ARTICLE INFO

Article history:
Received 30 April 2015
Revised 15 October 2015
Accepted 12 December 2015
Available online xxxx

Keywords:
Type 2 diabetes
Evidence-based medicine
Phenotype
Electronic health records
Population screening

ABSTRACT

Objectives: An estimated 25% of type two diabetes mellitus (DM2) patients in the United States are undiagnosed due to inadequate screening, because it is prohibitive to administer laboratory tests to everyone. We assess whether electronic health record (EHR) phenotyping could improve DM2 screening compared to conventional models, even when records are incomplete and not recorded systematically across patients and practice locations, as is typically seen in practice.

Methods: In this cross-sectional, retrospective study, EHR data from 9948 US patients were used to develop a pre-screening tool to predict current DM2, using multivariate logistic regression and a random-forests probabilistic model for out-of-sample validation. We compared (1) a full EHR model containing commonly prescribed medications, diagnoses, and conventional predictors, (2) a restricted EHR DX model which excluded medications, and (3) a conventional model containing basic predictors and their interactions (BMI, age, sex, smoking status, hypertension).

Results: Using a patient's full EHR or restricted EHR was superior to using basic covariates alone for detecting individuals with diabetes (hierarchical X^2 test, $p < 0.001$). Migraines and cardiac dysrhythmias were associated negatively with DM2, while sexual and gender identity disorder diagnosis and herpes zoster were associated positively. Adding EHR phenotypes improved classification; the AUC for the full EHR Model, EHR DX model, and conventional model using logistic regression, were 84.9%, 83.2%, and 75.0% respectively. For random forest out-of-sample prediction, accuracy also was improved when using EHR phenotypes; the AUC values were 81.3%, 79.6%, and 74.8%, respectively. Improved AUCs reflect better performance for most thresholds that balance sensitivity and specificity.

Conclusions: EHR phenotyping resulted in markedly superior detection of DM2, even in the face of missing and unsystematically recorded data, based on the ROC curves. EHR phenotypes could more efficiently identify which patients do require, and don't require, further laboratory screening. When applied to the current number of undiagnosed individuals in the United States, we predict that incorporating EHR phenotype screening would identify an additional 400,000 patients with active, untreated diabetes compared to the conventional pre-screening models.

1. Introduction

Although roughly 25% of people with type 2 diabetes mellitus (DM2) are undiagnosed in the United States, population-wide screening for diabetes currently is not cost-effective, because of the additional time and laboratory testing required [1]. Intervention studies have shown that diabetes can be prevented in high-risk individuals [1], while weight loss and lifestyle changes can revert the recently diagnosed patients (<4 years) to pre-diabetic state [2]; this makes population-wide screening not just an issue of prevention, but also one of treatment.

The total estimated cost of diagnosed diabetes in 2012 reached a staggering $245 billion, a 41% increase since 2007. People with diagnosed diabetes, on average, have medical expenditures

* Corresponding author at: Semel Institute, 760 Westwood Plaza, Ste B8-169, Los Angeles, CA 90095-1406, , United States. Tel.: +1 (310) 254 5680.
E-mail address: WesleyTK@UCLA.edu (W.T. Kerr).

http://dx.doi.org/10.1016/j.jbi.2015.12.006
1532-0464/

approximately 2.3 times higher than people who do not [3]. Characterizing diabetes risk using electronic health records (EHR), as used routinely for billing, could better estimate the financial cost of covering and treating an at-risk population. In this way, EHRs could extend screening models, conventionally framed between the doctor and the patient, to a predictive model between the payer and the patient. This could encourage targeted patient-incentive and education programs for at-risk populations.

Currently, comprehensive diabetes screening risk scores combine basic demographic and historical information with laboratory testing, to predict the future likelihood of developing diabetes. Laboratory tests can include fasting plasma glucose concentration, oral glucose tolerance test, or hemoglobin A1c (compared more thoroughly in [4]). These tests often require fasting, patient monitoring and blood draws, which can place an unmanageable burden on the patients, staff, and treating physicians when applied on the scale of millions of patients. This is particularly problematic in the resource limited health-care settings which are the most likely to service at-risk patients [5,6].

Diabetes screening is recommended by the U.S. Preventive Services Task Force only for asymptomatic adults with treated or untreated blood pressure over 135/80 mmHg, even though hypertension is only one of many known risk factors for diabetes [7]. In our sample, this would miss 1 in 4 patients diagnosed with DM2, while unnecessarily screening 1 in 3 patients without a recorded DM2 diagnosis. These data suggest that more sophisticated screening methods are needed, consistent with the Wilson and Jungner criteria [8,9].

While EHRs have demonstrated potential for detecting and monitoring diabetes [1], previous studies have used only a subset of all information available in the medical record, and typically have assessed risk only on patients for whom there were specific laboratory results available (e.g., fasting plasma glucose). EHR-based phenotypes can identify individuals who may benefit from interventions and thereby improve patient treatment and prognosis [10,11]. For example, usage of an EHR was associated with a decreased rate of emergency department visits in individuals with diabetes [1], and EHR data have been used to compute the prospective risk of developing dementia in individuals with diabetes [12].

If realistic results are desired data mining methods should be validated against real-world data. Records of "typical" quality are missing large amounts of data, with unsystematic data collection and recordings across practice locations. We examine whether augmenting risk scores using EHR-derived phenotypes would increase the ability to detect patients who should be screened further using laboratory testing, even when records are incomplete, and are not recorded systematically across health professionals and/or practice locations. When implemented on a population, this step-wise screening process would decrease the public health cost of more expensive testing, while simultaneously identifying previously overlooked at-risk patients.

2. Subjects

The study population included approximately 131,000 unique EHR transcript (visit) entries, containing 9948 patients from 1137 unique sites spanning all 50 United States, collected between 2009 and 2012, supplied in https://www.kaggle.com/c/pf2012-diabetes/data. Table 1 contains further demographic information. DM2 was diagnosed in 18.1% of patients according to at least one corresponding diagnosis within ICD9 250.X category (no patients had mixed Type 1/Type 2 diagnoses). We use the term "unrecorded" to describe patients without a DM2 diagnosis rather than the term "healthy", because the patients without a recorded DM2 diagnosis had more prescribed medications, and higher smoking rates, than patients with diabetes mellitus. This dataset is public and de-identified, provided by the free web-based EHR company, Practice Fusion. We intentionally used an unselected patient population who had a wide variety of laboratory tests, prescribed medications, and diagnoses. This dataset was rich in the breadth of information it contained, but did not include the free-text notes written about each patient (see Supplemental Methods for list of included factors).

Table 1
Demographic and basic information about the patients included in the study.

Mean (standard deviation)	Unrecorded control	Type 1 diabetes	Type 2 diabetes
Number of Patients (n)	7978	165	1805
Male (%)	40.6%	51.5%	50.6%
Age (years)	51 (18)	56 (15)	63 (13)
BMI (kg/m^2)	29 (6)	29 (7)	29 (6)
Systolic BP (mm Hg)	126 (18)	128 (19)	127 (19)
Diastolic BP (mm Hg)	77 (11)	77 (12)	77 (11)
Total Medications Prescribed	4.5 (4.5)	4.0 (4.0)	4.3 (4.6)
Total Diabetes Risk Factors	0.7 (0.9)	1.1 (.9)	1.2 (.9)
Hypertension DX (%)	34.5%	64.2%	72.5%
High Cholesterol (%)	28.7%	51.5%	62.4%
Smoking (%)	6.3%	5.4%	5.4%

Unless otherwise specified, the dataset assumed patients were healthy, took no medications, and underwent no laboratory tests. Missing entries were not identified clearly; a patient who had no history of taking a medication may have used yet not reported it. Consequentially, less than 1% of patients reported a family history of diabetes (ICD9 V18.0), despite a prevalence of 11.8% in the US population. It is unknown whether patients identified as unrecorded DM2 actually had undiagnosed DM2, likely due to current screening guidelines. Therefore, the dataset underestimates the prevalence of most disorders. This posed a "worst case" scenario for prediction; given missing, unsystematic and incomplete information from a patient's medical history, could residual information still augment current diabetes risk scores in a way that improves the accuracy and efficiency of DM2 screening in the general population?

3. Materials and methods

We assessed whether DM2 risk scores could be improved with EHR phenotypes, created using the additional medical and diagnostic information contained in the EHR. Because the visit dates were removed to protect patient privacy, information from multiple visits was combined across the whole study period into one data point representing each patient. The absence of visit dates made us unable to determine whether patients developed diabetes during their time of service, or whether it preceded their entry into this study. Similarly, the temporal ordering of medications, non-diabetes diagnoses, and the diabetes diagnosis are similarly unknown. Using real-world clinical data, these models then assess the current likelihood of a patient having a current diagnosis of DM2, rather than the future likelihood of developing diabetes.

We predicted current DM2 status using a multivariate logistic regression in R [13] comparing three separate models: (1) conventional model mimicking conventional risk scores; (2) a full "EHR Model" based upon the EHR phenotype, containing conventional information and both diagnostic and prescription information; and (3) "EHR DX" model which contained conventional information along with selected EHR information, excluding only medications. Within the "EHR DX" model, prescription information was removed because a diabetes diagnosis could change which medications physicians would prescribe. A partial list of predictive factors is illustrated in Table 2.

Please cite this article in press as: A.E. Anderson et al., Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.12.006