Title

Developing the Quantitative Histopathology Image Ontology (QHIO): A case study using the hot spot detection problem

Authors

Gurcan, Metin N
Tomaszewski, John
Overton, James A
et al.

Publication Date

2017-02-01

DOI

10.1016/j.jbi.2016.12.006

Peer reviewed

Journal of Biomedical Informatics 60 (2016) 162–168

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 17 December 2015

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 (as ICD9 categories), 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, depot medroxyprogesterone acetate, and cardiac dysrhythmias were associated negatively with DM2, while sexual and gender identity disorder diagnosis, viral and chlamydial infections, 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 machine learning 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.

The first model (conventional model) mimics conventional risk scores by including only the limited subset of covariates (smoking status, sex, age, BMI, and hypertensive status) that have been used in current diabetes risk models [4,14], and included all interaction effects.

The second "EHR Model" used 298 features: 150 most common diagnoses, 150 most commonly prescribed medications (before condensing name-brand and generic), transcript information (Table 2), and other specialized features summarized in Table 2. Hypertensive status, hyperlipidemia, and metabolic diagnosis all were assessed separately. To reduce bias, we removed as predictors established treatments and complications of diabetes mellitus: primary and secondary diabetes-related diagnoses (ICD-9 250.X2, 249.X), foot ulcers (ICD-9 707.X), diabetic retinopathy (ICD-9 362.01), polyneuropathy in diabetes (ICD-9 357.2), diabetic cataract (ICD-9 366.41), and diabetes mellitus complicating pregnancy, childbirth or the puerperium (ICD-9 648.0X). Medications used to treat diabetes, such as metformin (Glucophage™), were excluded from the model.

For the EHR Model, we created an "additional risk factor" variable tallying common comorbidities of diabetes, including conditions that have been shown to be more common in diabetes, but may be caused by factors other than diabetes. These included: candidiasis of skin and nails, malignant neoplasm of pancreas, other disorders of pancreatic internal secretion, polycystic ovaries, disorders of lipoid metabolism, overweight, obesity, and other hyperalimentation, trigeminal nerve disorders, hypertensive heart disease, acute myocardial infarction, other acute and subacute forms of ischemic heart disease, old myocardial infarction, angina pectoris, other forms of chronic ischemic heart disease, atherosclerosis, gingival and periodontal diseases, disorders of menstruation and other abnormal bleeding from female genital tract, unspecified local infection of skin and subcutaneous tissue, acquired acanthosis nigricans and tachypnea (Supplementary Embedded Table 1).

The third "EHR DX Model" used all features as in the EHR model, yet excluded all the medications. This was based on the assumption that a clinician's prescribing behavior would likely be influenced by a patient's diabetes status. Given that this sample is not longitudinal, excluding medication information reduces the bias inherent in all observational studies.

The conventional model was used as a reference standard in lieu of established risk scores because it ensures that difference between the two models is attributable to the structure and covariates, instead of the underlying study populations. We compared these models in a hierarchical regression with a chi-square test, and computed receiver operating characteristic (ROC) curves by measuring the area under the curve (AUC) within the R package ROCR [15].

Within the EHR model, the significance of each covariate was evaluated using a Wald test which keeps the expected false positive rate below the 0.05 threshold. Briefly, a Wald test divides the magnitude of the estimated log odds ratio by the standard error of that estimate. This quotient is compared to a t-distribution, with degrees of freedom based on the number of independent patient samples, minus the number of covariates in the model. Because many factors were found to be statistically significant, we make our interpretation more conservative by also adjusting for the false discovery rate (FDR) using the graphically sharpened method, setting the maximum proportion of false discoveries at 5% [16,17] as implemented by Reed et al. [18]. The FDR is computed over all 298 estimated p-values. We indicate which variables had q-values (adjusted p-values) in the FDR significance range using an asterisk within the regression tables, along with odds-ratio 95%-confidence intervals which take into account the prevalence of the risk factors being considered.

Finally, we validated using EHR phenotypes externally using a random forests probabilistic (not binary) prediction model, to assess the sensitivity of these models to overfitting [19]. In the random forests model, decision trees are constructed by resampling with replacement from the data and the predictors. Resampling with replacement causes roughly 30% patients to not be included each decision tree. An unbiased estimate of validation accuracy was achieved by probabilistic prediction of individual cases using decision trees that did not sample that individual observation using the data from the EHR model, the EHR DX model, and the conventional model. The probabilistic random forest models are used to create ROC curves that can be compared to the predicted probability from the logistic models.

4. Results

Incorporating EHR phenotypes improved classification accuracy more greatly than using limited covariates, for both the logistic regression and the random forests models.

The Full EHR model, and the EHR DX models, both predicted better than the conventional models (X^2 test, $p < 0.001$, Fig. 1). For the Full EHR Model, EHR DX model, and conventional model,

Table 2
Summary of non-medication or ICD9 code EHR information used to create full and restricted EHR models. For list of medications and ICD9 codes utilized, see Supplemental Table 1 and 2.

Variable	Description
Demographic Information	Age and gender
Transcript information	Systolic/Diastolic BP, Height, Weight, BMI
Diagnosis	150 most common ICD9 Codes (primary digits, i.e. 152.X)
Medications	150 most commonly prescribed (before collapsing generics)
Hypertension Diagnosis	ICD9 401-405
Cholesterol Diagnosis	ICD9 272
Metabolic Diagnosis	ICD9 277.7
Diagnosis	Total number of diagnoses over 4 year period
Acute Diagnosis	Total number of acute diagnoses over 4 year period
Additional Risk Factors	Diagnostic comorbidities, but not complications, of diabetes
Family Metabolic Disorder History	ICD9 V18.1, V18.11, V18.19
Lab Results	Total number acquired, number abnormal and percent of total that were abnormal
Geographical Region (State#)	Defined by census.gov/geo/maps-data/maps/pdfs/reference/us_regdiv.pdf
State Diabetes Rate	2010 National Diabetes Surveillance System Location Diabetes Rate
Family Diabetes Risk	ICD9 V18.0
Family Metabolic Risk	ICD9 V18.1, V18.11, V18.19
Family Diabetes History	ICD9 V18.0
High Risk Practice (loc4)	Average diabetes rate is above 8.3% at that practice location
Smoking Status	Current, Former, Never Smoked, NA (NIST Codes)

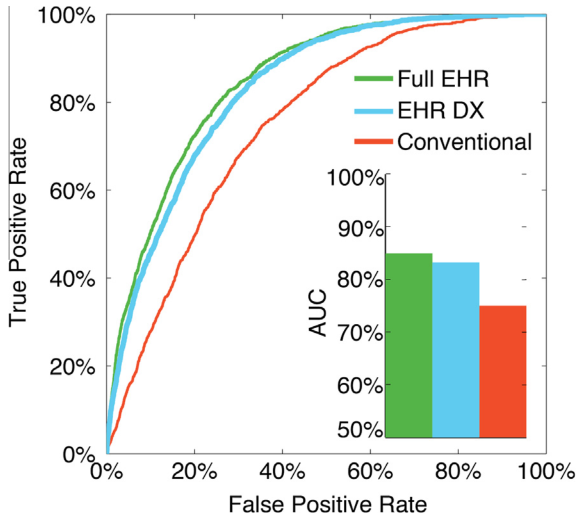

Fig. 1. Receiver Operating Curve (ROC) for each of the logistic models, and their respective area under the ROC (AUC). Supplemental Fig. 4 is the same figure for the random forest models. For all classification thresholds, incorporating EHR phenotypes improved the ability to detect patients with DM2.

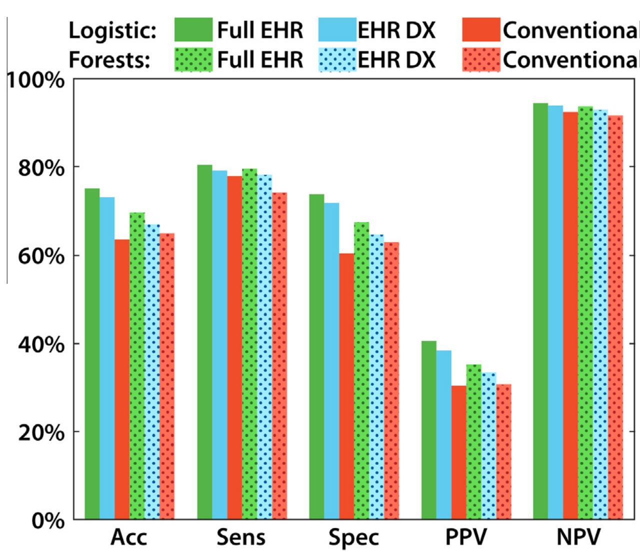

Fig. 2. The performance of each type of algorithm (logistic and random forest) for each type of input data (full EHR, EHR DX and conventional). The threshold used to assess performance corresponded to an 18.1% prevalence of diabetes in the full dataset. For all algorithms, incorporating more EHR data led to higher performance in identifying DM2 and non-DM2 patients, compared to conventional algorithms. Shaded bars indicate random forest models. Error bars were too small to display meaningfully with 9948 subjects. Significance markings indicate hierarchical X^2 tests for $p < 0.05$. Abbreviations: accuracy (acc), sensitivity (sens), specificity (spec), positive and negative predictive value (PPV and NPV).

the AUC was 84.9%, 83.2% and 75.0%, respectively (Fig. 2). The random forests AUC for the full EHR model, the EHR DX model, and the conventional model were 81.3%, 79.6%, and 74.8%, respectively (Supplemental Fig. 1). Significantly predictive factors for the full EHR model are illustrated in Fig. 3.

The accuracy, sensitivity, specificity, positive predictive values, and negative predictive values all depend on the decision threshold for the probabilistic classification. We set the threshold so that 18.1% would be diagnosed with diabetes, consistent with the prevalence of diagnosed diabetes in the full dataset. At this threshold, the EHR models had better performance in all categories, holding constant the model family (random forest vs. logistic). This is illustrated in Fig. 2, with 95% confidence intervals for all values provided in Supplemental Text.

In the conventional model the variables age, hypertensive status, and the interaction of age with hypertension, all were significant predictors of DM2 (Wald test, $p < 0.05$, Supplemental Fig. 2). The interaction between age and hypertensive status indicated that younger hypertensive patients had a greater chance of DM2 than elderly hypertensive patients. The 298 coefficients for the EHR DX Model, the full EHR model, and the conventional model are provided within the Supplementary Figs. 2, 3 and 4.

Within the held-out data (using the random forests probabilistic models), the EHR models had greater sensitivity than the conventional model for all thresholds, as shown in Fig. 2 and Supplemental Fig. 1. The overall accuracies were dependent on thresholds chosen, but for all thresholds, the ROC curve for the EHR models exceeded the sensitivity of the conventional model as shown in Supplementary Fig. 2.

5. Discussion

The EHR phenotype models outperformed the basic screening model for detecting individuals at-risk for diabetes, for all thresholds. While many of the risk factors identified in the full EHR model have been identified previously, they have not been evaluated within the context of all other clinical factors. Due to the structure of our multivariate logistic regression model, each discussed factor was significant, controlling for, and separate from, all other measured factors, while adjusting for the false discovery rate. Additionally, we demonstrated that the EHR had strong predictive power when excluding all physician-prescribed medications, suggesting that the diagnoses contain nearly the same information as the medications prescribed to treat them.

The number of studied and significant factors was too large to consider the effect of each individually, so we limit our discussion to selected salient factors which additionally surpassed the false discovery rate (for a more extended discussion of each factor, see Supplemental Discussion). These associations are based on patients diagnosed with DM2, who may or may not be similar to patients with unrecorded DM2. The list of factors positively associated with diabetes was consistent with previous literature and, in particular, included hyperlipidemia and hypertension – the other two pillars of metabolic syndrome – as strong predictive factors for DM2 [20]. The factors negatively associated with diabetes seemed to be indicators for physical activity and chronic disease that involved regular monitoring and integration into the medical system. In addition to these interpretable factors, our full EHR approach identified unexpected factors that were associated significantly with current DM2 including ICD9 302.X (sexual and gender identity disorders), ICD9 477.X (allergic rhinitis) and the use of depot medroxyprogesterone contraceptives.

One of the benefits of EHR research is that, in addition to verifying the factors that had a relatively clear interpretation, EHR mining can identify others where the link with disease is either unproven or unsuspected. Some factors identified here are less established in the DM2 literature. In particular, although some work has addressed the association of homosexuality, sexual identity disorders, and sexual deviancy (ICD-9 302.X) with diabetes [21,22], our results suggest that more work should be done to understand the link between diabetes and these factors, so that at-risk patients can be identified better. In contrast to increasing the risk of current DM2, we are uncertain why allergic rhinitis (ICD-9 477.X), and use of depot medroxyprogesterone contraceptives, decreased the prevalence of DM2. Even though DM2 has been shown to affect innate or acute immunity [23], we are unaware of a strong link between DM2 and allergies, which primarily

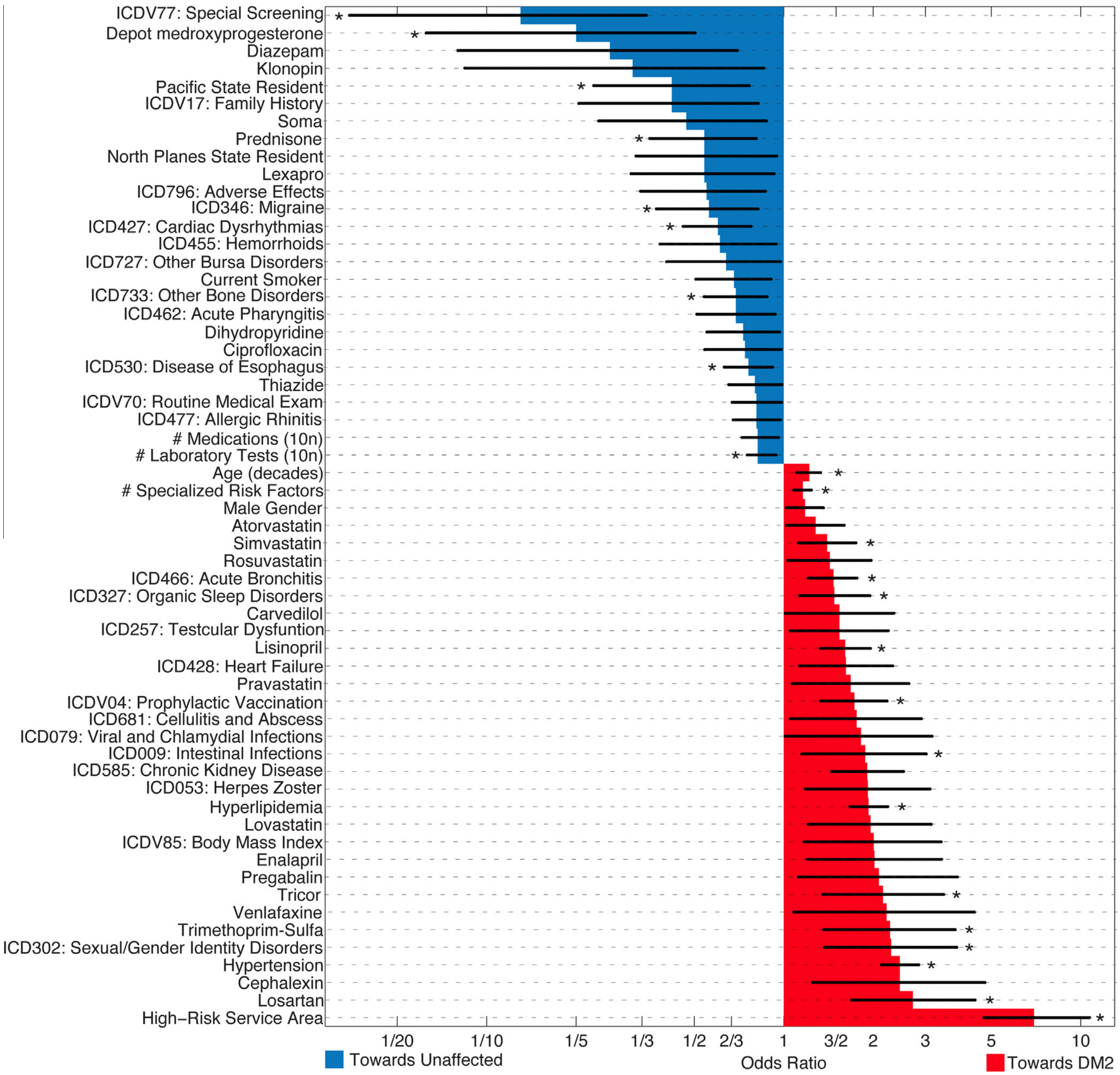

Fig. 3. Statistically significant EHR factors positively (red) and negatively (blue) associated with having a diabetes diagnosis. Star (*) indicates significance after additional false-discovery rate correction. For all modeled factors, see Supplemental Fig. 3. Error bars reflect 95% confidence intervals. See Table 2 and Supplemental Tables 1 and 2 for definitions of variables, definitions of the included ICD9 codes and medication categories.

are a dysfunction of the adaptive or acquired immune system. Due to the underlying etiology behind the previous controversial link between exogenous hormone treatment for menopause and cardiovascular risk, we expect that the usage of depot medroxyprogesterone contraceptives reflects a patient population that engages in other activities that decrease DM2 risk, and not that depot medroxyprogesterone itself is protective for DM2. The interpretation of each of these factors is unclear; therefore our results suggest that more work should be done to understand these observed links.

Some unexpected factors positively associated with diabetes include hemorrhoids (ICD-9 455), medications used to treat anxiety disorders and seizures (diazepam, clonazepam), and disease of the esophagus (ICD-9 530). Unexpected negative factors included viral and chlamydial infections (ICD-9 079), organic sleep disorders (ICD-9 327), and intestinal infections (ICD9-009).

In addition to these factors, there were a number of features that were associated negatively with DM2, even though the literature suggested that the association should be positive. An established side effect of prednisone treatment is increased insulin resistance and steroid-induced DM2 [24]. Therefore, clinicians may prescribe prednisone only in low-risk patients. As noted above, DM2 is associated positively with disorders of the innate or acute immune system, metabolic syndrome and cardiac dysfunction; therefore we are uncertain why some of these factors were associated negatively with DM2. Migraine also shares common comorbidities with DM2 [25], including some that we identified here. Our EHR model also found that being a current smoker decreased the risk of diabetes, which reflects how smoking increases base metabolic rate. Yet, this contradicts at least one study in the literature [26] which found smoking increased diabetes risk. However, this difference may be due to how we controlled for many other factors, while previous studies on smoking and diabetes accounted for a limited number of confounders (e.g., age & BMI). These factors warrant further study to understand why our analysis of EHR records did not replicate previous work.

There are several limitations to our model. Our model faced challenges with incomplete data such as family history, and did not contain several important risk factors for diabetes, such as ethnicity and socioeconomic status [27], due to incomplete patient records. Given that the prevalence of diabetes in the sample population (18.1%) greatly exceeded the general US diabetes prevalence (11.8%), this suggests that our sample population may reflect the population of Americans with health insurance treated during the 2.5 year study period. Certainly, patients with diabetes or symptoms of diabetes are more likely to have insurance and seek medical attention. Alternatively, given the prevalence of unrecorded DM2, the clinics that contributed to Practice Fusion may be better at identifying DM2 than the average clinic. The incomplete nature of the records data affected mainly the full EHR model, as the basic covariates (age, BMI, etc.) were not missing. If incomplete data had been indicated more systematically, missing data could have been imputed to improve accuracy and reduce bias. Although we established a protocol for interpreting medication and laboratory test results, there was large variation in the reporting of these factors; therefore some bias and/or misreporting could be present. Due to privacy concerns, our model was unable to incorporate longitudinal information in the EHR that we would expect to improve its overall accuracy. Moreover, this model was trained using patients who had a current diagnosis of DM2, which implicitly assumes that diagnosed and undiagnosed diabetes patients are similar, but this hypothesis needs to be confirmed. In the future, we will confirm the profiles of diagnosed and undiagnosed patients using longitudinal data on an independent database, with a 12–18 month pre-index period with no diagnoses for Type I or II. Finally, the covariates used to predict DM2 were all necessarily statistically dependent, so the significant factors may be blocking the true impact of other correlated factors.

Our model is not prognostic; the factors we have found associated with DM2 are not causal. However, given the finding that DM2 was diagnosed most frequently after a patient exhibited at least one complication, and that 25% of patients are unrecorded [28], both early prediction and identification of untreated DM2 are critical to clinical practice. We specifically excluded known complications of diabetes to reduce this bias, but the complications and risk factors of diabetes often can be intertwined (e.g., cardiovascular disease).

Given that this analysis showed superior prediction of diabetes using only 3 years of incomplete and unsystematically recorded data, we anticipate that the true signal and potential of an ideal, complete, and systematically utilized EHR is much greater than we have demonstrated here. As EHRs become used more widely, we anticipate that the size and quality of the records will increase by orders of magnitude. Increase in sample size, and reduction of the amount of missing data, would only strengthen the ability of these models to detect DM2 and reliably identify at-risk patients.

We, and others, have advocated that data mining EHRs could be used to address numerous clinical challenges [1], such as identifying patients at risk for depression, suicide, strokes, and cardiovascular events. Risk scores, not limited just to diabetes, could be automatically computed and included in a general patient profile, providing physicians an instant assessment of potential health conditions.

Beyond prediction, using EHR-phenotype models provide invaluable information about the risk factors themselves. For example, EHRs can be used to assess comparative risks and benefits of medication classes (as is being done [29]), and answer important treatment questions, including whether statins or fibrates are more effective to treat high cholesterol in patients with diabetes. However, our experience with hormonal therapy for menopause [30] taught us that, while there is great potential in retrospective, observational studies, the highest level of evidence is a double-blinded, randomized clinical trial. Because of this, resulting associations with diabetes are necessarily ambiguous, with no defined causal relationship, and need to be scrutinized carefully in light of complementary controlled studies [31]. For screening of a disease that is both the fifth leading cause of preventable death in the United States, and that has a tremendous rate of unrecorded patients, we argue that the accuracy of the model is more important than any causal inference of the predictors.

6. Conclusion

Given that nearly 1 in 3 Americans will develop diabetes at some point in their lifetime, predicting and assessing diabetes risk on a population-wide scale is critical for both prevention and effective treatment. As a positive consequence of the U.S. and U.K. EHR mandates, an EHR phenotype-based pre-screen could be used for national diabetes screening at little cost, as risk scores could be computed automatically within EHRs to efficiently identify at-risk patients who should undergo more formal and sensitive laboratory screening and/or preventive behavioral interventions.

We anticipate that the largest adopter of EHR phenotype models will be public and private insurers; risk scores could be created using existing claims databases. Given that the average patient with diabetes incurs over twice the expected cost as a patient without [3] and that the Affordable Care Act of 2010 bars insurers from dropping coverage of ill patients or denying coverage because of pre-existing conditions, there exists a significant financial advantage for insurers to incentivize their high-risk patients. Individual patients who are at high-risk for chronic and costly diseases could be targeted for patient-education programs, and reduction in calculated risk would be rewarded monetarily.

These promising results showed that EHR phenotypes provided superior predictive accuracy for assessing diabetes status, compared to traditional non-laboratory information ($p < 0.001$). We have demonstrated that this is possible even in the face of a diverse, at-risk patient population, with missing and incomplete patient records pooled across practice locations. This suggests that incorporating more medical history could increase the accuracy of existing diabetes risk scores in at-risk patient populations, for stepwise screening. The combined efficacy of EHR screening plus focused laboratory testing needs future study.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Ginger Haynes for her contributions to this manuscript. We acknowledge NIH R33DA026109 to M.S.C., the University of California (UCLA)-California Institute of Technology Medical Scientist Training Program (NIH T32 GM08042), the Systems and Integrative Biology Training Program at UCLA (NIH T32 GM008185), and the UCLA Department of Biomathematics for partial funding of this research. We additionally thank Practice Fusion, Inc. for providing data publicly for research. Ariana E. Anderson, Ph. D., holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.12.006.

References

[1] Control CfD, Prevention, Control CfD and Prevention, National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States, 2011, US Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA, 2011, p. 201.

[2] R. Kahn, P. Alperin, D. Eddy, et al., Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis, Lancet 375 (2010) 1365–1374.

[3] C.L. Gillies, P.C. Lambert, K.R. Abrams, et al., Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis, Bmj 336 (2008) 1180–1185.

[4] J. Tuomilehto, J. Lindström, J.G. Eriksson, et al., Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance, New England J. Med. 344 (2001) 1343–1350.

[5] X.-R. Pan, G.-w. Li, Y.-H. Hu, et al., Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. the Da Qing IGT and Diabetes Study, Diabetes Care 20 (1997) 537–544.

[6] E. Lim, K. Hollingsworth, B. Aribisala, M. Chen, J. Mathers, R. Taylor, Reversal of type 2 diabetes: normalisation of beta cell function in association with decreased pancreas and liver triacylglycerol, Diabetologia 54 (2011) 2506–2514.

[7] A.D. Association, Economic costs of diabetes in the US in 2012, Diabetes Care 36 (2013) 1033–1046.

[8] P. Schwarz, J. Li, J. Lindstrom, J. Tuomilehto, Tools for predicting the risk of type 2 diabetes in daily practice, Horm. Metab. Res. 41 (2009) 86–97.

[9] N. Calonge, D.B. Petitti, T.G. DeWitt, et al., Screening for type 2 diabetes mellitus in adults: US Preventive Services Task Force recommendation statement, Ann. Internal Med. 148 (2008) 846–854.

[10] J. Wilson, G. Jungner, Principles and Practice of Screening for Disease, World Health Organization, Geneva, 1968, Public Health Papers, 2011, p. 34.

[11] A.M. Sheehy, D.B. Coursin, R.A. Gabbay, Back to Wilson and Jungner: 10 good reasons to screen for type 2 diabetes mellitus, in: Mayo Clinic Proceedings, Mayo Foundation, 2009, p. 38.

[12] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from EMR data using machine learning, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2012, p. 606.

[13] S. Griffin, P. Little, C. Hales, A. Kinmonth, N. Wareham, Diabetes risk score: towards earlier detection of type 2 diabetes in general practice, Diabetes/Metab. Res. Rev. 16 (2000) 164–171.

[14] E.P.K. Woolthuis, W.J. de Grauw, W.H. van Gerwen, et al., Identifying people at risk for undiagnosed type 2 diabetes using the GP's electronic medical record, Family Pract. 24 (2007) 230–236.

[15] J. Hippisley-Cox, C. Coupland, J. Robson, A. Sheikh, P. Brindle, Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore, BMJ: British Med. J. (2009) 338.

[16] K.M. Newton, P.L. Peissig, A.N. Kho, et al., Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, J. Am. Med. Inf. Assoc. 20 (2013) e147–e154.

[17] R.D. Cebul, T.E. Love, A.K. Jain, C.J. Hebert, Electronic health records and quality of diabetes care, New England J. Med. 365 (2011) 825–833.

[18] M. Reed, J. Huang, R. Brand, et al., Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes, JAMA 310 (2013) 1060–1065.

[19] S.M. Speedie, Y.-T. Park, J. Du, et al., The impact of electronic health records on people with diabetes in three different emergency departments, J. Am. Med. Inf. Assoc. 21 (2014) e71–e77.

[20] L.G. Exalto, G.J. Biessels, A.J. Karter, et al., Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study, Lancet Diabetes Endoc. 1 (2013) 183–190.

[21] Team RDC, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2005.

[22] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics 21 (2005) 3940–3941.

[23] Y. Benjamini, Y. Hochberg, On the adaptive control of the false discovery rate in multiple testing with independent statistics, J. Educ. Behav. Stat. 25 (2000) 60–83.

[24] K.J. Verhoeven, K.L. Simonsen, L.M. McIntyre, Implementing false discovery rate control: increasing your power, Oikos 108 (2005) 643–647.

[25] N. Pike, Using false discovery rates for multiple comparisons in ecology and evolution, Methods Ecol. Evol. 2 (2011) 278–282.

[26] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[27] S.M. Grundy, H.B. Brewer, J.I. Cleeman, S.C. Smith, C. Lenfant, Definition of metabolic syndrome report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on scientific issues related to definition, Circulation 109 (2004) 433–438.

[28] M.I. Harris, Undiagnosed NIDDM: clinical and public health issues, Diabetes Care 16 (1993) 642–652.

[29] F.S. Roque, P.B. Jensen, H. Schmock, et al., Using electronic patient records to discover disease correlations and stratify patient cohorts, PLoS Comput. Biol. 7 (2011) e1002141.

[30] M. Masica, M. Collinsworth, Leveraging electronic health records in comparative effectiveness research, Prescriptions Excellence Health Care Newsletter Suppl. 1 (2012) 6.

[31] W.T. Kerr, E.P. Lau, G.E. Owens, A. Trefler, The future of medical diagnostics: large digitized databases, Yale J. Biol. Med. 85 (2012) 363.