Big Data
Volume 3 Number 4, 2015

DOI: 10.1089/big.2015.0020

ORIGINAL ARTICLE

Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors

Narges Razavian,[1] Saul Blecker,[2] Ann Marie Schmidt,[3] Aaron Smith-McLallen,[4] Somesh Nigam,[4] and David Sontag[1,*]

Abstract

We present a new approach to population health, in which data-driven predictive models are learned for outcomes such as type 2 diabetes. Our approach enables risk assessment from readily available electronic claims data on large populations, without additional screening cost. Proposed model uncovers early and late-stage risk factors. Using administrative claims, pharmacy records, healthcare utilization, and laboratory results of 4.1 million individuals between 2005 and 2009, an initial set of 42,000 variables were derived that together describe the full health status and history of every individual. Machine learning was then used to methodically enhance predictive variable set and fit models predicting onset of type 2 diabetes in 2009–2011, 2010–2012, and 2011–2013. We compared the enhanced model with a parsimonious model consisting of known diabetes risk factors in a real-world environment, where missing values are common and prevalent. Furthermore, we analyzed novel and known risk factors emerging from the model at different age groups at different stages before the onset. Parsimonious model using 21 classic diabetes risk factors resulted in area under ROC curve (AUC) of 0.75 for diabetes prediction within a 2-year window following the baseline. The enhanced model increased the AUC to 0.80, with about 900 variables selected as predictive ($p < 0.0001$ for differences between AUCs). Similar improvements were observed for models predicting diabetes onset 1–3 years and 2–4 years after baseline. The enhanced model improved positive predictive value by at least 50% and identified novel surrogate risk factors for type 2 diabetes, such as chronic liver disease (odds ratio [OR] 3.71), high alanine aminotransferase (OR 2.26), esophageal reflux (OR 1.85), and history of acute bronchitis (OR 1.45). Liver risk factors emerge later in the process of diabetes development compared with obesity-related factors such as hypertension and high hemoglobin A1c. In conclusion, population-level risk prediction for type 2 diabetes using readily available administrative data is feasible and has better prediction performance than classical diabetes risk prediction algorithms on very large populations with missing data. The new model enables intervention allocation at national scale quickly and accurately and recovers potentially novel risk factors at different stages before the disease onset.

Key words: big data analytics; data mining; machine learning; predictive analytics; risk assessment; disease prediction; longitudinal study

Introduction

The recent availability of the electronic health record and claims datasets offers an unprecedented opportunity to apply predictive analytics to improve the practice of medicine and to infer potentially novel risk factors.[1–3] Successful examples of previously deployed large-scale risk assessment models include hospital readmission models,[4,5] disease onset prediction,[6–13] and prediction of healthcare utilization and cost.[14]

Type 2 diabetes is a global public health challenge. The total number of people with diabetes (including type 1 and 2) is estimated to rise from 171 million in 2000 to 366 million in 2030,[15] and current statistics show that the vast majority of diabetic patients are suffering from type 2 diabetes.[16] In 2002, the Centers for Disease Control (CDC) Diabetes Prevention Program (DPP)[17] showed that intensive lifestyle intervention[18] focusing on exercise and weight loss was more effective

[1]*Department of Computer Science, New York University, New York, New York.*
[2]*Department of Population Health, NYU Langone Medical Center, New York University, New York, New York.*
[3]*Department of Medicine, Department of Biochemistry and Molecular Pharmacology, Department of Pathology Medicine, and Diabetes Research Program, NYU Langone Medical Center, New York University, New York, New York.*
[4]*Advanced Analytics, Independence Blue Cross, Philadelphia, Pennsylvania.*

*Address correspondence to: *David Sontag, Department of Computer Science, New York University, 715 Broadway, 12th Floor, Room 1204, New York, NY 10003,* E-mail: dsontag@cs.nyu.edu

at lowering the risk of type 2 diabetes than medication with Metformin. Similar studies in other countries have confirmed the benefits.[19–21]

Despite the academic success of the DPP, implementation of the program by major insurance or public health service providers has been hindered by a number of limitations. The interventions can only be cost-effective when the target population has a high likelihood of developing diabetes at the baseline.[22] The DPP program, which selected participants based on obesity and elevated glucose levels, observed only an 11% positive predictive value (PPV)[18] (i.e., only 11% of the participants without any lifestyle or Metformin intervention developed diabetes within 3 years). This emphasizes the need for models with better risk assessment for diabetes onset. Traditional well-known models for type 2 diabetes onset, including ARIC,[23] San-Antonio,[24] AUSDRISK,[25] and FINDRISC,[26] provide potential solutions for more accurate risk assessment, but these models have another major limitation: they require a time-consuming and costly screening step, which again makes the interventions infeasible.

The primary purpose of our study is to develop a population-level risk prediction model for type 2 diabetes, which can be directly applied to health insurance claims and other readily available clinical and utilization data. Using machine learning, we methodically discover surrogates for variables that would otherwise be missing.

The secondary purpose of our study is to identify the relative importance of different risk factors in terms of how early they may predict onset of type 2 diabetes. Observational studies using clinical and utilization data provide a window into the lives of patients before clinical diagnosis of type 2 diabetes at a scale much larger than what would be feasible within the scope of a clinical trial or prospective cohort study.

Materials and Methods

We performed a retrospective cohort study of beneficiaries of Independence Blue Cross (Independence), a major insurance provider in southeastern Pennsylvania. The primary data source for the study was Independence claims data, which included enrollment information, utilization records such as hospitalizations, outpatient visits, laboratory orders, and medication fulfillment, for all beneficiaries, and laboratory test results for 95% of the laboratory claims.

Our initial population included ~4.1 million deidentified Independence beneficiaries, at least 18 years of age, who enrolled in Independence's insurance program between the years 2005 and 2013.

Outcome

Our primary outcome was the confirmed diagnosis of type 2 diabetes. A beneficiary was confirmed as having type 2 diabetes if any of the following three criteria were observed on two distinct days: (1) an International Classification of Diseases, Clinical Modification (ICD-9-CM) code of 250.xx, listed as a hospital discharge diagnosis or physician clinical encounter; (2) use of a diabetes medication, including Glimepiride, Glipizide, Glyburide, Chlorpropamide, Tolazamide, Tolbutamide, Pioglitazone, Rosiglitazone, Acarbose, Miglitol, Repaglinide, Nateglinide, Sitagliptin, Saxagliptin, Linagliptin, Alogliptin, Pramlintide, Exenatide, Liraglutide, Canagliflozin, and Insulin (any); or (3) HbA1C value ≥6.5%. The list of medications was based on existing diabetes outcome definitions,[27] and we excluded Metformin from the definition of diabetes due to its significant usage in treatment of polycystic ovarian syndrome[28] and prediabetes. To derive our final outcome definition, we compared the performance of multiple clinically relevant definitions of diabetes among a representative subgroup of beneficiaries who, based on the Standard of Care for Diabetes,[29] could be definitively labeled as either having diabetes or being free from diabetes (see Supplementary Material, Part-A; Supplementary Data are available online at www.liebertpub.com/big).

Parsimonious model based on known risk factors: baseline

We built a parsimonious baseline model, using risk factors derived from seven landmark studies of risk prediction models for predicting incident diabetes: ARIC,[23] KORA,[30] FRAMINGHAM,[31] AUSDRISC,[25] FINDRISC,[26] and the San-Antonio Model.[24] To build our parsimonious model, we included every variable that was used in any of these models for which we had direct or surrogate measurements. The variables include age (continuous variable), gender (binary indicator), overweight (binary indicator), underweight (binary indicator), diagnosis of obesity (binary indicator), hypercholesterolemia history (binary indicator), cardiovascular disease history (binary indicator), lipid disorder history (binary indicator), history of high alcohol in blood (binary indicator), history of unspecified hypertension (binary indicator), prediabetic fasting glucose level (binary variable, set to 1 if the fasting glucose

level was ≥100.0 and ≤125.0), high triglyceride level (binary variable, set to 1 if the triglyceride level was ≥150.0), high C-reactive protein level (binary variable, set to 1 if level ≥0.75 percentile),[32] and a protective HDL-C level (binary variable, set to 1 if value ≥40 for male or ≥50 for female). We included the diagnosis of obesity as a surrogate variable for high body–mass index (BMI), and the diagnoses of hypertension and hypertensive heart and renal diseases as surrogates for elevated blood pressure. Because the cited models were not developed for use with claims datasets, and moreover in some cases we used surrogates for variables not observable in claims data, we retrained the parameters of the parsimonious baseline using the training data.

Enhanced model

We next built an enhanced model using beneficiary demographics (11 continuous and binary variables), including age as one continuous variable in addition to three binary variables for age intervals of 18–39, 40–64, and 65+ years, gender, and number of months with vision and dental insurance coverage; all past and current medical conditions (16,632 binary variables); temporal procedures received (457 variables for each of three different time intervals); temporal physician specialty visits (50×3 binary variables); temporal laboratory orders and results (7000×3 binary variables); and temporal medication utilization (990×3 binary variables). For all temporal variables, we calculated the feature over the past 6 months, 24 months, and entire past history. If a variable was not observed, it was referenced as 0 and we did not impute it.

Medical conditions were encoded as indicator variables based on all International Classification of Diseases (ICD-9) diagnosis codes. In our initial studies, we had used the Clinical Classification Software (CCS)[33] hierarchy of ICD-9 codes in a temporal manner. However, we saw no gain in the predictive power compared with using individual diagnosis variables. To preserve the granularity of the risk factors, we therefore did not encode past medical conditions temporally or hierarchically. Procedure information variables were based on the Current Procedure Terminology (CPT) and ICD-9 procedural codes, each grouped by CCS.[33] Additional variables included indicators for visiting every physician specialty possible in clinical encounters (which are available in claims data) and indicators for all medications as specified by the national drug code and grouped by therapeutic class codes.

Patient laboratory measurement variables were based on logical observation identifiers, names, and codes. We used the 1000 most frequent laboratory tests based on our cohort. For each of these laboratory tests at each time span considered, we derived seven variables: an indicator of whether the test was administered, an indicator for whether the result was reported as low, high, or normal according to the reference range of the laboratory, and whether the value increased, decreased, or fluctuated.

In total, each beneficiary was represented as a set of $\sim$42,000 variables that summarized all their past and current medical states. We emphasize that these variables were not selected specifically for the purpose of studying type 2 diabetes. Our approach thus has the potential to discover novel risk factors associated with type 2 diabetes.

Study framework and inclusion criteria

We designed our study to determine the risk of developing type 2 diabetes in three time spans into the future. We built the feature vectors from beneficiary data up to December 31, 2008. We then predicted whether subjects were to develop type 2 diabetes within a 2-year prediction window: immediately within the following 2 years (i.e., between January 1, 2009, and January 1, 2011); at least 1 year into the future (i.e., between January 1, 2010, and January 1, 2012); or at least 2 years into the future (i.e., between January 1, 2011, and January 1, 2013). For each of these analyses, we excluded beneficiaries who had already developed diabetes before the start of the prediction window. We required a minimum of 6 months of enrollment before December 31, 2008, to include the beneficiary in our study. The framework is summarized in Figure 1. We refer to the time span between data collection and the beginning of the prediction window as the gap period.

Statistical analysis

We developed the prediction models using sparse, or L1-regularized, logistic regression.[34] This method provides a computationally efficient alternative to commonly used variable selection methods, such as forward selection and backward elimination, and eliminates both variable ordering bias and the need to adjust for the p-value inflation coming from multiple comparison tests on the same dataset.[35] L1 regularization simultaneously searches over all relevant and irrelevant variables.[36] It comes with a strong mathematical guarantee to recover the true set of predictors and learn the corresponding

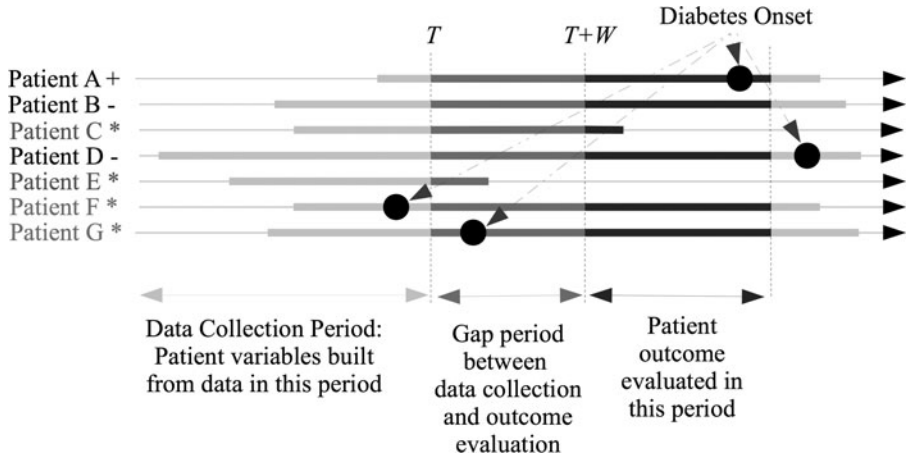

FIG. 1. Framework for the prediction task. Features are derived from patient data up to time *T*. Outcome is evaluated in the 2-year follow-up window after a gap of size *W*. Patients who have diabetes before *T*+*W*, or have insufficient enrollment, are excluded during training and evaluation (denoted as *). Patient outcome is positive (denoted as +) if diabetes onset happens in the outcome evaluation period and negative (denoted as −) otherwise.

beta coefficients, even when the number of samples is smaller than the number of irrelevant variables.[37]

L1 regularization works by adding a penalty to the classification loss. This penalty is the sum of absolute values of the coefficients (called L1 penalty) and guides the optimization algorithm to select a beta coefficient vector that pushes very low weights to zero when those low weights do not improve the accuracy of the prediction. As a result, the final beta coefficient vectors will be sparse, interpretable, robust to noise, and statistically powerful. Fast algorithms to optimize the accuracy of such models are available.[36,38] We use an algorithm based on Dual Coordinate Descent,[38] which handles massive datasets very efficiently, to train these models from data.[39,40] We optimized a reweighted log likelihood to correct for the class imbalance during the training. We use the area under the ROC curve (AUC) as the primary evaluation metric. AUC is invariant to the prior class probabilities and thus suitable for imbalanced datasets.

We used randomly selected 67% of the data for training, with the remaining 33% held out for the validation set, and used a fivefold cross-validation on the training data to choose the L1 regularization hyperparameter. For regularization parameters, we searched over values of [0.001, 0.01, 0.1, 1, 10], and 0.1 was selected to be optimum in all our settings and cross-validation folds. We used the same methodology and the reweighted log likelihood objective to fit the parameters of the parsimonious model. Additional details of our method for presenting the results are included in Supplementary Part-B.

For each predictive model, we calculated the AUC on the validation set. We also report a PPV for the 100, 1000, and 10,000 individuals predicted to develop diabetes with highest probability based on our enhanced models, compared with the parsimonious model, using the validation data. We calculated the odds ratio (OR) for each discovered risk factor and present them for three age categories. In all cases, we report the unadjusted ORs directly calculated from the data, which link each risk factor to diabetes onset independently of the other variables. For all reported risk factors, we reported 95% confidence intervals (CIs) in addition to *p*-values for the ORs. To report AUC CIs, we used a standard error upper bound[41] and reported 95% CIs. For PPV, we reported 95% CI.[42] In all comparisons, we used the Wald test for reporting *p*-values of differences.

Results

Data

The original cohort included about 4.1 million beneficiaries. A total of 793,153 beneficiaries matched the inclusion criteria for predicting onset of type 2 diabetes between January 1, 2009, and January 1, 2011, using beneficiaries' data through December 31, 2008. These

Table 1. Subjects' characteristics of the cohort included in training and validation

Characteristic	*Total population*	*Population with diabetes*
Average age (SD)	47.69 (17.1)	58.57 (13.3)
Female ratio	55%	51%
Average length of data in years (SD)	3.3 (1.0)	3.4 (1.0)
Hypertension (ICD9 401)	30.2%	62%
Hypercholesterolemia (ICD9 272.0)	18.7%	33.6%

SD, standard deviation.

beneficiaries' characteristics are included in Table 1. Of these, 19,307 developed diabetes within the prediction window. After training, 967 variables were selected for the enhanced model. For predicting onset of type 2 diabetes between January 1, 2010, and January 1, 2012 (gap period 1 year), a total of 697,502 beneficiaries matched the inclusion criteria; of these 13,835 beneficiaries developed diabetes within the prediction window. After training, 769 variables were selected in the enhanced model as predictive. For predicting onset of type 2 diabetes between January 1, 2011, and January 1, 2013, 629,817 beneficiaries matched our inclusion criteria, 8498 of which had a positive label in the prediction window. After training, 538 variables were selected as predictive.

Prediction results

For immediate prediction of diabetes, the enhanced model had an AUC of 0.80 compared with an AUC of 0.75 for the baseline parsimonious model ($p < 0.0001$). The AUCs of the enhanced models were also superior to those of the parsimonious models for prediction of diabetes in the 1- or 2-year gap periods (Table 2). Similarly, PPV values for the top 100, 1000, and 10,000 beneficiaries predicted to be diabetic were between 1.6 and 2.3 times higher in the enhanced model than the parsimonious model (Table 2). Our models are highly specific, and the sensitivity increases to 21% at the 10,000 level. The ROC curves corresponding to the enhanced versus parsimonious models for different gap periods are included in Figure 2. Predicting onset of diabetes further into the future, with a larger gap between data collection and the evaluation window, is (expectedly) less accurate.

Table 3 shows the top predictive variables for immediate onset of diabetes. Most top variables are directly related to prediabetes or diabetes, including history of prediabetes or related conditions, elevated glucose, elevated HbA1c, and Metformin medication utilization. However, other variables, such as history of sleep apnea, acute bronchitis, hypothyroidism, and anemia, as well as high serum alanine aminotransferase, have significant predictive value for immediate confirmation of onset of diabetes. Measures of healthcare utilization also contribute to the prediction of onset of type 2 diabetes. Expanded lists of the laboratory values and disease history that are predictive of diabetes diagnosis are included in Supplementary Tables S1 and S2. Noteworthy is the difference in the OR within the young, middle-aged, and older population for different factors. Specifically, the OR of all risk factors is higher when the factor is observed in younger individuals. OR of factors such as elevated HbA1c and glucose in the young population is almost thrice that of the middle-aged population, and almost four times that of the older population.

Table 4 shows the top predictive variables for diabetes onset within 1 to 3 years after the data collection period (gap = 1). Not surprisingly, previously identified risk factors such as high glucose, high HbA1c, obesity, and impaired fasting glucose emerged as strongly predictive of diabetes diagnosis. Interestingly, 1 year

Table 2. Performance for prediction of diabetes, using patient data through December 31, 2008, within the different prediction windows

			Top 100[b]			*Top 1000*[b]			*Top 10,000*[b]		
Prediction window	*Model*	*AUC*[a,b]	*Sensitivity*	*Specificity*	*PPV*	*Sensitivity*	*Specificity*	*PPV*	*Sensitivity*	*Specificity*	*PPV*
2009–2011	Parsimonious model	0.75	0.001	0.999	0.12	0.014	0.996	0.10	0.114	0.967	0.08
	Enhanced model	0.80	0.005	0.999	0.37	0.033	0.997	0.23	0.216	0.969	0.15
2010–2012	Parsimonious model	0.74	0.001	0.999	0.06	0.014	0.996	0.07	0.117	0.962	0.06
	Enhanced model	0.78	0.002	0.999	0.15	0.035	0.996	0.17	0.203	0.963	0.10
2011–2013	Parsimonious model	0.72	0.0009	0.999	0.03	0.012	0.995	0.04	0.118	0.957	0.03
	Enhanced model	0.76	0.003	0.999	0.10	0.024	0.995	0.07	0.195	0.958	0.06

[a]Differences in AUC significant with $p < 0.0001$ in this validation set.
[b]All reported values have 95% CI of less than 0.002.
AUC, area under curve; CI, confidence interval; PPV, positive predictive value.

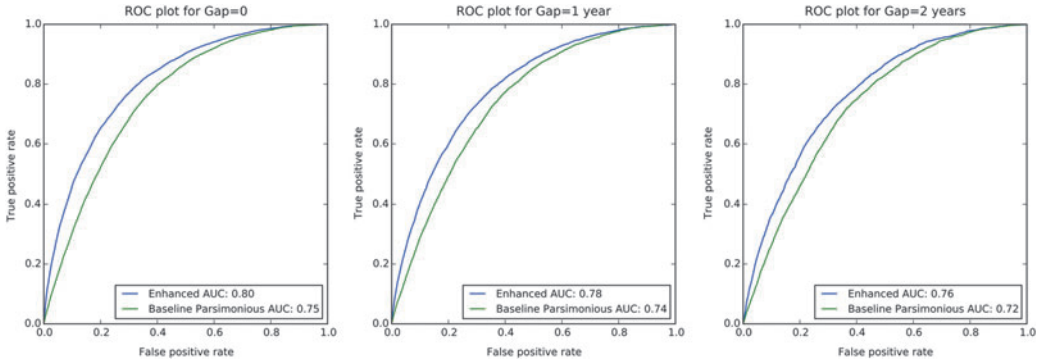

FIG. 2. ROC curves for predicting type 2 diabetes onset for 0, 1, and 2 years into the future.

before the confirmed diagnosis of diabetes, shortness of breath, esophageal reflux, and acute bronchitis also have significant predictive value. Healthcare usage variables such as need for emergency room service and routine child health examination are also significant in assessment of risk of impending diabetes. Expanded lists of predictive laboratory values and disease history are included in Supplementary Tables S3 and S4. The top predictive variables for the model with the 2-year gap are included in Supplementary Table S5.

Discussion

Related work on machine learning and data mining for early detection of type 2 diabetes can be categorized into three groups. The first group of related works uses the classic clinical diabetes risk prediction studies,[23–26,30] which focus on large cohorts, but rely on small feature sets and logistic regression models. Our parsimonious baseline is based on these models. The second group uses classical diabetes risk factors as features, but focuses on comparing machine learning models such as Support Vector Machines,[43] CART,[44] and decision trees[45] to combine the features.[46–49] These models do not consider the potential impact of using a broader set of features. In contrast, our work uses a broad, rich, set of features together with a linear model trained with L1-regularized logistic regression. Although it is beyond the scope of the current study to explore in detail, we found that this was better than or comparable with carefully tuned random forests,[50] gradient-boosted decision trees,[51] and neural networks[52] on our dataset.

The third group of related work considers a broader set of features, which can be utilized to predict outcomes such as heart failure[12,53] and the occurrence of urgent care events and uncontrolled A1c for diabetic patients.[6] These related works also use (variants of) logistic regression for predictive modeling. Our approach is related to this group in terms of the generalizability of the method to multiple outcomes. However, we focus on diabetes onset as the outcome and provide an in-depth analysis of the selected features. Moreover, we investigate differences in risk factors at multiple stages before the disease onset. The temporal aspect of the risk factors, both in terms of early or late risk factors,[46] and also the temporal trends for variables such as laboratory measurements[54,55] are less studied. Temporal features are often studied for a handful of variables at a time, whereas in our study, we utilize basic temporal patterns on all 1000 laboratory measurements.

The present study is the largest study to date of early detection of type 2 diabetes, both in terms of cohort size and the number of variables considered. Model fitting and validation were conducted using more than 42,000 variables. Hundreds of variables were selected as predictive of future type 2 diabetes. We demonstrated that compared with using a parsimonious set of variables, using big data and machine learning improves PPVs by 67% and AUC by 6.6%. The resulting models are already deployed at Independence Blue Cross for the purpose of intervention allocation.

Our risk models do not require additional[56] tests, screening, or chart reviews beyond what is readily available in health records and administrative data. This allows our approach to scale to millions of beneficiaries on a regular schedule. The reported sensitivity, specificity, and positive predictive values for our models can provide guidance for intervention targeting. For focused high-cost interventions, our method is able to

Table 3. Top predictive variables for type 2 diabetes onset within 2009–2010 (gap = 0), using patient data through December 31, 2008

Variable type	*Variable evaluation period*[a]	*Variable description*	*Number with diabetes*	*Number without diabetes*	*OR (95% CI)*	*OR for 18 ≤ age <40*	*OR for 40 ≤ age <65*	*OR for 65 ≤ age*	*p-value of OR*
Laboratory test	Past 2 years	Hemoglobin A1c/hemoglobin.total—high (LOINC-4548-4)	1845	8710	9.28 (8.81 9.78)	23.01 (16.8 31.40)	8.42 (7.85 9.03)	4.34 (4.00 4.72)	<0.001
	Past 2 years	Glucose—high (LOINC-2345-7)	5274	58,736	4.58 (4.43 4.73)	9.42 (7.90 11.24)	3.68 (3.52 3.84)	2.42 (2.29 2.56)	<0.001
	Past 2 years	Hemoglobin A1c/hemoglobin.total—request for test (LOINC-4548-4)	3908	45,519	4.06 (3.92 4.21)	5.90 (5.03 6.91)	3.41 (3.25 3.57)	2.56 (2.40 2.73)	<0.001
	Entire history	Cholesterol.in HDL—low (LOINC-2085-9)	3233	49,524	2.94 (2.83 3.06)	4.72 (3.99 5.59)	2.41 (2.29 2.53)	1.99 (1.86 2.14)	<0.001
	Entire history	Triglyceride—high (LOINC-2571-8)	6056	106,818	2.85 (2.77 2.94)	3.92 (3.37 4.55)	2.29 (2.20 2.38)	1.64 (1.55 1.73)	<0.001
	Entire history	Cholesterol.total/cholesterol.in HDL—high (LOINC-9830-1)	3114	56,032	2.46 (2.37 2.56)	4.12 (3.41 4.99)	2.04 (1.94 2.14)	1.47 (1.37 1.58)	<0.001
	Entire history	Alanine aminotransferase—high (LOINC-1742-6)	1208	22,205	2.26 (2.13 2.40)	3.49 (2.74 4.46)	2.00 (1.86 2.15)	1.53 (1.37 1.72)	<0.001
	Entire history	Cholesterol.in VLDL—request for test (LOINC-13458-5)	3029	63,166	2.09 (2.01 2.18)	2.60 (2.16 3.14)	1.67 (1.59 1.76)	1.54 (1.44 1.65)	<0.001
	Entire history	Cholesterol.total/cholesterol.in HDL—decreasing (LOINC-9830-1)	3277	75,701	1.89 (1.81 1.96)	2.76 (2.15 3.55)	1.40 (1.33 1.48)	1.07 (1.01 1.14)	<0.001
	Past 2 years	Carbon dioxide—request for test (LOINC-2028-9)	6044	158,472	1.77 (1.72 1.83)	2.59 (2.25 2.98)	1.28 (1.23 1.34)	1.12 (1.06 1.18)	<0.001
ICD9 history	Entire history	Abnormal glucose (ICD9 790.29)	1198	10,099	5.00 (4.70 5.32)	10.64 (7.89 14.35)	4.31 (3.98 4.68)	2.64 (2.39 2.92)	<0.001
	Entire history	Impaired fasting glucose (ICD9 790.21)	1285	11,521	4.72 (4.45 5.01)	9.82 (6.69 14.41)	4.04 (3.74 4.37)	2.38 (2.16 2.62)	<0.001
	Entire history	Hypertension (ICD9 401)	12,175	227,759	4.09 (3.97 4.22)	4.77 (4.21 5.41)	2.94 (2.84 3.05)	1.95 (1.83 2.09)	<0.001
	Entire history	Chronic liver disease (ICD9 571.8)	619	6845	3.71 (3.41 4.03)	7.46 (5.22 10.66)	3.32 (3.01 3.66)	2.00 (1.68 2.39)	<0.001
	Entire history	Obesity (ICD9 278)	3104	48,000	2.90 (2.78 3.01)	4.71 (4.10 5.40)	2.85 (2.71 2.98)	1.97 (1.81 2.14)	<0.001
	Entire history	Obstructive sleep apnea (ICD9 327.23)	1178	17,302	2.84 (2.67 3.02)	4.11 (3.07 5.50)	2.48 (2.30 2.66)	1.81 (1.60 2.05)	<0.001
	Entire history	Hypersomnia with sleep apnea (ICD9 780.53)	1138	16,965	2.79 (2.63 2.97)	4.15 (3.04 5.67)	2.38 (2.21 2.56)	1.83 (1.62 2.08)	<0.001
	Entire history	Abnormal blood chemistry (ICD9 790.6)	2388	38,726	2.68 (2.56 2.80)	3.54 (2.83 4.43)	2.28 (2.15 2.41)	1.57 (1.46 1.69)	<0.001
	Entire history	Hyperlipidemia (ICD9 272.4	8745	186,016	2.62 (2.54 2.69)	3.31 (2.87 3.82)	1.86 (1.79 1.93)	1.40 (1.33 1.48)	<0.001
	Entire history	Anemia (ICD9 285.9)	3421	75,500	1.99 (1.92 2.07)	2.74 (2.34 3.20)	1.63 (1.55 1.72)	1.39 (1.31 1.48)	<0.001
	Entire history	Hypothyroidism (ICD9 244.9)	3803	87,228	1.93 (1.86 2.00)	3.35 (2.85 3.93)	1.53 (1.46 1.60)	1.17 (1.10 1.25)	<0.001
	Entire history	Acute bronchitis (ICD9 466.0)	3229	93,559	1.46 (1.41 1.52)	1.64 (1.40 1.92)	1.30 (1.24 1.37)	1.20 (1.12 1.29)	<0.001
NDC medication history	Entire history	Medication group: Metformin	286	1142	10.17 (8.93 11.59)	17.17 (12.67 23.25)	11.38 (9.57 13.53)	12.76 (9.36 17.39)	<0.001
	Entire history	Medication group: antiarthritics	3055	88,506	1.46 (1.40 1.51)	1.74 (1.49 2.03)	1.25 (1.19 1.32)	1.22 (1.14 1.31)	<0.001
	Entire history	Medication group: nonsteroidal anti-inflammatory drugs	3216	94,531	1.44 (1.38 1.49)	1.72 (1.47 2.00)	1.24 (1.18 1.30)	1.22 (1.14 1.31)	<0.001
Healthcare utilization	Past 2 years	Procedure group: routine chest X	5505	131,707	1.94 (1.88 2.01)	1.86 (1.61 2.15)	1.60 (1.53 1.67)	1.30 (1.23 1.37)	<0.001
	Entire history	Service place code: home	4386	113,223	1.72 (1.66 1.77)	1.69 (1.47 1.94)	1.56 (1.49 1.63)	1.26 (1.19 1.34)	<0.001
	Entire history	Dental coverage = yes	4142	119,108	1.50 (1.45 1.55)	1.04 (0.89 1.23)	1.05 (0.99 1.11)	1.17 (1.11 1.24)	<0.001
	Entire history	Specialty code: internal medicine	7246	227,156	1.45 (1.40 1.49)	1.64 (1.46 1.86)	1.12 (1.08 1.16)	1.00 (0.95 1.05)	<0.001
	Entire history	Procedure group: ophthalmologic and otologic diagnosis and treatment	6681	247,300	1.13 (1.09 1.16)	0.86 (0.76 0.97)	1.04 (1.00 1.08)	0.89 (0.84 0.93)	<0.001

Shown here are the variables with the highest magnitude of beta coefficient, sorted by the unadjusted OR.
[a]Entire history refers to our current setting and cohort, which is limited to max 4 years before 2009.
OR, odds ratio.

Table 4. Top predictive variables for type 2 diabetes onset within 2010–2012 (gap = 1), using patient data through December 31, 2008

Variable type	*Variable evaluation period*[a]	*Variable description*	*Number with diabetes*	*Number without diabetes*	*OR (95% CI)*	*OR for 18 ≤ age <40*	*OR for 40 ≤ age <65*	*OR for 65 ≤ age*	p-*value of OR*
Laboratory test	Entire history	Hemoglobin A1c/hemoglobin.total—high (LOINC-4548-4)	1323	12,344	5.75 (5.42 6.10)	7.98 (5.58 11.41)	5.46 (5.05 5.90)	2.74 (2.49 3.02)	<0.001
	Past 2 years	Glucose—high (LOINC-2345-7)	3389	50,745	4.05 (3.89 4.21)	7.31 (5.88 9.10)	3.24 (3.07 3.41)	2.25 (2.10 2.40)	<0.001
	Past 2 years	Hemoglobin A1c/hemoglobin.total—request for test	2389	39,347	3.42 (3.27 3.58)	5.11 (4.23 6.17)	2.90 (2.74 3.07)	2.14 (1.97 2.32)	<0.001
	Entire history	Hemoglobin A1c/hemoglobin.total—request for test	3111	58,061	3.13 (3.00 3.26)	4.63 (3.91 5.47)	2.63 (2.49 2.77)	1.94 (1.81 2.09)	<0.001
	Entire history	Cholesterol.In HDL—low (LOINC-2085-9)	2172	42,888	2.78 (2.66 2.92)	4.69 (3.88 5.68)	2.27 (2.14 2.41)	1.90 (1.75 2.08)	<0.001
	Entire history	Cholesterol.total/cholesterol.In HDL—high (LOINC-9830-1)	2082	49,026	2.29 (2.19 2.40)	4.00 (3.22 4.97)	1.87 (1.76 1.98)	1.42 (1.30 1.55)	<0.001
	Entire history	Cholesterol.In VLDL—request for test (LOINC-13458-5)	2277	55,592	2.23 (2.13 2.33)	2.43 (1.96 3.01)	1.80 (1.70 1.91)	1.67 (1.54 1.81)	<0.001
	Entire history	Carbon dioxide—request for test (LOINC-2028-9)	5157	186,669	1.58 (1.53 1.64)	2.58 (2.24 2.96)	1.13 (1.08 1.18)	0.99 (0.93 1.06)	<0.001
	Past 2 years	Glomerular filtration rate/1.73 Sq. M.Predicted.Black—request for test (LOINC-48643-1)	3560	123,104	1.58 (1.52 1.64)	2.37 (2.00 2.81)	1.15 (1.09 1.21)	1.04 (0.97 1.11)	<0.001
ICD9 history	Entire history	Impaired fasting glucose (ICD9-790.21)	800	9918	4.17 (3.87 4.49)	7.05 (4.27 11.65)	3.42 (3.10 3.77)	2.33 (2.07 2.62)	<0.001
	Entire history	Abnormal glucose NEC (ICD9-790.29)	690	8695	4.07 (3.76 4.41)	7.46 (5.05 11.00)	3.46 (3.12 3.84)	2.28 (2.01 2.60)	<0.001
	Entire history	Hypertension (ICD9-401)	6026	130,309	3.28 (3.17 3.39)	4.60 (3.88 5.44)	2.55 (2.44 2.66)	1.64 (1.53 1.75)	<0.001
	Entire history	Obstructive sleep apnea (ICD9-327.23)	867	14,979	2.98 (2.78 3.20)	4.50 (3.30 6.15)	2.61 (2.40 2.84)	1.89 (1.63 2.19)	<0.001
	Entire history	Obesity (ICD9 278)	2189	41,850	2.88 (2.75 3.02)	4.44 (3.80 5.19)	2.81 (2.66 2.97)	2.01 (1.81 2.22)	<0.001
	Entire history	Abnormal blood chemistry (ICD9-790.6)	1588	33,877	2.49 (2.36 2.62)	3.81 (2.99 4.86)	2.08 (1.94 2.23)	1.51 (1.38 1.65)	<0.001
	Entire history	Hyperlipidemia (ICD9 272.4)	6017	163,558	2.45 (2.37 2.53)	3.09 (2.63 3.65)	1.74 (1.66 1.81)	1.40 (1.31 1.50)	<0.001
	Entire history	Shortness of breath (ICD9-786.05)	2132	54,848	2.09 (1.99 2.19)	2.23 (1.80 2.76)	1.78 (1.67 1.89)	1.38 (1.28 1.50)	<0.001
	Entire history	Esophageal reflux (ICD9-530.81)	2889	85,302	1.85 (1.78 1.93)	2.12 (1.75 2.56)	1.52 (1.44 1.60)	1.23 (1.14 1.32)	<0.001
	Entire history	Acute bronchitis (ICD9-466.0)	2273	82,255	1.44 (1.37 1.50)	1.49 (1.24 1.78)	1.30 (1.22 1.37)	1.20 (1.11 1.31)	<0.001
NDC medications	Past 2 years	Medication group: antiarthritics	2109	76,497	1.43 (1.36 1.50)	1.67 (1.40 2.00)	1.22 (1.15 1.29)	1.22 (1.12 1.33)	<0.001
	Entire history	Medication group: antiarthritics	2230	81,802	1.41 (1.35 1.48)	1.68 (1.41 2.00)	1.20 (1.14 1.28)	1.23 (1.13 1.33)	<0.001
Healthcare utilization	Entire history	Procedure group: routine chest X-ray	4973	152,365	1.96 (1.89 2.03)	2.05 (1.78 2.36)	1.58 (1.51 1.66)	1.33 (1.24 1.41)	<0.001
	Entire history	Dental coverage = yes	2919	105,445	1.47 (1.41 1.53)	1.08 (0.90 1.30)	1.08 (1.01 1.16)	1.14 (1.07 1.22)	<0.001
	Entire history	Service place: emergency room—hospital	5920	246,865	1.32 (1.28 1.37)	1.39 (1.23 1.56)	1.41 (1.35 1.47)	1.29 (1.21 1.37)	<0.001
	Entire history	Specialty code: independent laboratory	6946	314,429	1.18 (1.14 1.22)	1.28 (1.13 1.44)	0.98 (0.94 1.02)	1.01 (0.95 1.08)	<0.001
	Entire history	Routine medical examination (ICD9 V700)	3432	191,452	0.85 (0.82 0.88)	1.06 (0.92 1.22)	0.76 (0.72 0.79)	0.75 (0.70 0.81)	<0.001
	Entire history	Routine gynecological examination (ICD9 V7231)	4448	246,649	0.84 (0.81 0.87)	1.75 (1.55 1.97)	0.69 (0.66 0.72)	0.86 (0.80 0.92)	<0.001
	Entire history	Routine child health examination (ICD9 V202)	175	76,181	0.10 (0.09 0.12)	0.31 (0.26 0.36)	0.41 (0.29 0.58)	0.39 (0.05 2.82)	<0.001

Shown here are the variables with the highest magnitude of beta coefficient, sorted by the unadjusted OR.
[a]Entire history refers to our current setting and cohort, which is limited to max 4 years before 2009.
LOINC, logical observation identifiers, names, and codes; NDC, national drug code.

(with 37% positive predictive value compared with only 12% using the parsimonious baseline) select the most vulnerable. When the interventions are more scalable, they could be performed on the 10,000 most vulnerable individuals, with a sensitivity of 21.6% in a validation set of more than 220,000 beneficiaries compared with only 11.4% using traditional risk prediction methods.

Our study is strengthened by a derived definition for diabetes, which was found to have a highly accurate definition of diabetes in our dataset. This is an important strength as prior studies of predictive modeling for diabetes have primarily relied on screening tests and clinical evaluation on specific clinical visits.[23–26]

Our models include both known risk factors for diabetes and less established risk factors, many of which are likely surrogates for established risk factors. Many of the strongest predictors in our models describe prediabetes or elements of the metabolic syndrome, including elevated HbA1c, high blood sugar, and hyperlipidemia (Tables 3 and 4). While obesity was found to be predictive of diabetes in our model (Tables 3 and 4), it was likely underreported in the insurance claims as only 6% of beneficiaries were documented as obese, despite 35% of the American population being defined as obese according to the CDC.[57] As a result, some less traditional risk factors found in our models may be acting as partial surrogates for obesity and its risk on the development of diabetes. For instance, esophageal reflux, which is documented for 12.6% of our population (Table 4), is known to have a high prevalence in obesity and may act as a surrogate for obesity in our data.[58,59] Similarly, sleep apnea and markers of inflammation such as leukocytosis have known associations with metabolic syndrome (Tables 3 and 4).[60] We also found other makers of cardiovascular disease, such as coronary atherosclerosis (Supplementary Table S2), hypertension (Tables 3 and 4), and kidney disease (Table 4 and Supplementary Tables S1–S3), to be predictive of diabetes onset. Although these conditions may be complications of diabetes, they also may be risk factors or diagnosed before diabetes onset.[61,62]

A number of risk factors related to liver disease were included in the predictive model for diabetes, including elevated alanine aminotransferase (Table 3 and Supplementary Tables S1 and S3) and the presence of chronic liver disease (Table 3 and Supplementary Tables S2 and S4). Elevated liver function tests are early manifestations of insulin resistance[63] and are detectable earlier than fasting glycemia.[64] Nonalcoholic fatty liver disease (Supplementary Tables S2 and S4) has a well-documented association with both obesity and diabetes.[65] Our method also selects hypothyroidism (Table 3 and Supplementary Table S4), which has known causal effects for insulin resistance,[66] as predictive. Our model also included a number of cardiopulmonary findings, including acute bronchitis (Tables 3 and 4), shortness of breath (Table 4), and chest X-rays (Tables 3 and 4), as predictive for diabetes. While we are unable to make any direct associated link, these factors may reflect association between cardiovascular diseases such as heart failure with subsequent diabetes or may point to an association that individuals with acute pulmonary conditions may be at risk for the development of diabetes.

Our study has several limitations. First, there may be more missing data among beneficiaries who have only recently enrolled in the health insurance plan or who have little healthcare utilization, reducing the sensitivity of the model among these beneficiaries. A possible solution would be to complement the administrative data with data gathered by other sources, such as by mobile health applications. Second, the study population may not be representative of the whole of the United States as 80% of the studied population resides in the greater Philadelphia, which may contribute both demographic and behavioral bias. However, we emphasize that our models can be easily retrained with other insurance companies' or providers' data. Third, since our outcome is derived from clinical and utilization data, we are unable to determine if a person has existing, but undiagnosed and untreated, type 2 diabetes. Due to the lack of a true gold standard for diabetes in our population, we were unable to confirm the sensitivity of our diabetes definition. Fourth, our parsimonious model used obesity as a surrogate measure for BMI. We found that the obesity diagnosis was likely underreported in our dataset, which may have limited the accuracy of the parsimonious model. Nonetheless, such limitations are common in claims datasets and highlight why previous models may not always be practical for population-level risk assessment.

Conclusion

Machine learning on administrative data provides a powerful new tool for population health and clinical hypothesis generation for risk factor discovery, enabling population-level risk assessment that may help guide interventions to the most at-risk population. Using the approach described herein, it is possible to identify patients likely to develop type 2 diabetes with

at least 67% better PPV compared with traditional risk assessment methods for 0–2 years into the future. The extensive set of risk factors recovered by our method, for different stages of disease onset, can be a basis for additional hypothesis testing in medical research laboratories. Finally, our approach is general enough to be applied to different outcomes of interest, to build predictive models for different years into the future, and to analyze the risk factors as they emerge at different stages before the onset.

Acknowledgments

The authors thank Ravi Chawla for comments on the manuscript and Rahul Krishnan and Youngduck Choi for technical assistance. D.S. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Disclosure Statement

Research was financially supported by a grant from Independence Blue Cross, which also contributed the data for the study. The sponsor collected the data, reviewed the manuscript, and approved the decision to submit the manuscript for publication. All authors contributed to the conception and design of the study. N.R., S.B., A.M.S., and D.S. interpreted the data and performed the statistical analysis. N.R. drafted the manuscript, and all authors performed critical revision of the manuscript. Two authors, A.S.-M. and S.N., are employees at Independence Blue Cross. There are no other conflicts of interest to report.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nat Rev Genet 2012; 13:395–405.
2. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Med Care 2010; 48:S106–S113.
3. Roden DM, Xu H, Denny JC, Wilke RA. Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. Clin Pharmacol Ther 2012; 91:1083–1086.
4. Krumholz HM. Post-hospital syndrome—An acquired, transient condition of generalized risk. N Engl J Med 2013; 368:100–102.
5. Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care 2013; 51:368–373.
6. Neuvirth H, Ozery-Flato M, Hu J, et al. Toward personalized care management of patients at risk: The diabetes case study. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM, 2011. pp. 395–403.
7. Lloyd-Jones DM, Leip EP, Larson MG, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 2006; 113:791–798.
8. Maguire J, Dhar V. Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: Data-driven predictive analytics in healthcare. Health Syst 2013; 2:73–92.
9. Perotte A, Ranganath R, Hirsch JS, et al. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015; 22:872–880.
10. Letham B, Rudin C, McCormick TH, Madigan D. Building interpretable classifiers with rules using Bayesian analysis. Technical Report No. 609, Department of Statistics, University of Washington, 2012.
11. Wiens J, Campbell WN, Franklin ES, et al. Learning data-driven patient risk stratification models for *Clostridium difficile*. Open Forum Infect Dis 2014; 1:ofu045.
12. Sun J, Hu J, Luo D, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012; 2012:901–910.
13. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015; 7:299ra122.
14. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014; 33:1123–1131.
15. Wild S, Roglic G, Green A, et al. Global prevalence of diabetes estimates for the year 2000 and projections for 2030. Diabetes Care 2004; 27:1047–1053.
16. Centers for Disease Control Prevention. National diabetes statistics report: Estimates of diabetes and its burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services 2014.
17. Diabetes Prevention Program (DPP) Research Group. The Diabetes Prevention Program (DPP) description of lifestyle intervention. Diabetes Care 2002; 25:2165–2171.
18. Knowler WC, Barrett-Connor E, Fowler SE. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002; 346:393–403.
19. Lindström J, Louheranta A, Mannelin M, et al. The Finnish Diabetes Prevention Study (DPS): Lifestyle intervention and 3-year results on diet and physical activity. Diabetes Care 2003; 26:3230–3236.
20. Li G, Zhang P, Wang J, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: A 20-year follow-up study. Lancet 2008; 371:1783–1789.
21. Ramachandran A, Snehalatha C, Mary S, et al. The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). Diabetologia 2006; 49:289–297.
22. Hernan WH, Brandle M, Zhang P, et al. Costs associated with the primary prevention of type 2 diabetes mellitus in the diabetes prevention program. Diabetes Care 2003; 26:36–47.
23. Kahn HS, Cheng YJ, Thompson TJ, et al. Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years. Ann Intern Med 2009; 150:741–751.
24. Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Ann Intern Med 2002; 136:575–581.
25. Chen L, Magliano DJ, Balkau B, et al. AUSDRISK: An Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust 2010; 192:197–202.
26. Lindström J, Tuomilehto J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. Diabetes Care 2003; 26:725–731.
27. Chang H-Y, Weiner JP, Richards TM, et al. Predicting costs with diabetes complications severity index in claims data. Am J Manag Care 2012; 18:213–219.
28. Lord JM, Flight IHK, Norman RJ. Metformin in polycystic ovary syndrome: Systematic review and meta-analysis. BMJ 2003; 327:951–953.
29. American Diabetes Association. Standards of medical care in diabetes—2015 Abridged for primary care providers. Clin Diabetes 2015; 33: 97–111.
30. Rathmann W, Kowall B, Heier M, et al. Prediction models for incident Type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. Diabet Med 2010; 27:1116–1123.
31. Wilson PW, Meigs JB, Sullivan L, et al. Prediction of incident diabetes mellitus in middle-aged adults: The Framingham Offspring Study. Arch Intern Med 2007; 167:1068–1074.

32. King DE, Mainous AG, Buchanan TA, Pearson WS. C-reactive protein and glycemic control in adults with diabetes. Diabetes Care 2003; 26:1535–1539.
33. Agency for Healthcare Research and Quality. HCUP clinical classifications software for services and procedures. Healthcare Cost and Utilization Project (HCUP). Rockville, MD 2014.
34. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Methodol 1996; 58:267–288.
35. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test 2003; 12:1–77.
36. Shi J, Yin W, Osher S, Sajda P. A fast hybrid algorithm for large-scale l 1-regularized logistic regression. J Mach Learn Res 2010; 11:713–741.
37. Ng AY. On feature selection: Learning with exponentially many irrelevant features as training examples. In: Proceedings of the International Conference on Machine Learning, San Francisco, CA: Morgan Kaufmann Publishers Inc., 1998. pp. 404–412.
38. Yu H-F, Huang F-L, Lin C-J. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach Learn 2011; 85:41–75.
39. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011; 12:2825–2830.
40. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Med 2011; 9:103.
41. Mohri C. Confidence intervals for the area under the ROC curve. In: Advances in Neural Information Processing Systems. Curran Associates, 2005, p. 305.
42. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med 1998; 17:857–872.
43. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20:273–297.
44. Loh WY. Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov 2011; 1:14–23.
45. Quinlan JR. Induction of decision trees. Mach Learn 1986; 1:81–106.
46. Mani S, Chen Y, Elasy T, et al. Type 2 diabetes risk forecasting from EMR data using machine learning. AMIA Annu Symp Proc 2012; 2012:606–615.
47. Simon GJ, Schrom J, Castro MR, et al. Survival association rule mining towards type 2 diabetes risk assessment. AMIA Annu Symp Proc 2013; 2013:1293–1302.
48. Meng X-H, Huang Y-X, Rao D-P, et al. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 2013; 29:93–99.
49. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med 2002; 26:37–54.
50. Ho TK. Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition. IEEE, 1995. pp. 278–282.
51. Mason L, Baxter J, Bartlett PL, Frean MR. Boosting algorithms as gradient descent. In: Advances in Neural Information Processing Systems. MIT Press, 2000, pp. 512–518.
52. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436–444.
53. Wang F, Zhang P, Qian B, et al. Clinical risk prediction with multilinear sparse logistic regression. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2014. pp. 145–154.
54. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. AMIA Annu Symp Proc 2009; 2009:452–456.
55. Silvent AS, Dojat M, Garbay C. Multi-level temporal abstraction for medical scenario construction. Int J Adapt Control Signal Processing 2005; 19:377–394.
56. Kulzer B, Hermanns N, Gorges D, et al. Prevention of diabetes self-management program (PREDIAS): Effects on weight, metabolic risk factors, and behavioral outcomes. Diabetes Care 2009; 32:1143–1146.
57. National Center for Health Statistics. Health, United States, 2012: With special feature on emergency care. Hyattsville, MD, 2013. Available at: www.cdc.gov/nchs/data/hus/hus12.pdf
58. Murray L, Johnston B, Lane A, et al. Relationship between body mass and gastro-oesophageal reflux symptoms: The Bristol Helicobacter Project. Int J Epidemiol 2003; 32:645–650.
59. Fisher BL, Pennathur A, Mutnick JLM, Little AG. Obesity correlates with gastroesophageal reflux. Dig Dis Sci 1999; 44:2290–2294.
60. Ip MSM, Lam B, Ng MMT, Lam WK, et al. Obstructive sleep apnea is independently associated with insulin resistance. Am J Respir Crit Care Med 2002; 165:670–676.
61. D'Agostino RB, Hamman RF, Karter AJ, et al. Cardiovascular disease risk factors predict the development of type 2 Diabetes: The insulin resistance atherosclerosis study. Diabetes Care 2004; 27:2234–2240.
62. Hu FB, Stampfer MJ, Haffner SM, et al. Elevated risk of cardiovascular disease prior to clinical diagnosis of type 2 diabetes. Diabetes Care 2002; 25:1129–1134.
63. Lewis GF, Carpentier A, Adeli K, Giacca A. Disordered fat storage and mobilization in the pathogenesis of insulin resistance and type 2 diabetes. Endocr Rev 2002; 23:201–229.
64. Harris EH. Elevated liver function tests in type 2 diabetes. Clin Diabetes 2005; 23:115–119.
65. Marchesini G, Brizi M, Morselli-Labate AM, et al. Association of nonalcoholic fatty liver disease with insulin resistance. Am J Med 1999; 107:450–455.
66. Wang C. The relationship between type 2 diabetes mellitus and related thyroid diseases. J Diabetes Res 2013;2013:1–9.

Cite this article as: Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D (2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 3:4, 277–287, DOI: 10.1089/big.2015.0020.

Abbreviations Used

BMI = body–mass index
CCS = Clinical Classification Software
CDC = Centers for Disease Control
CI = confidence intervals
DPP = Diabetes Prevention Program
ICD-9 = International Classification of Diseases
LOINC = logical observation identifiers, names, and codes
NDC = national drug code
OR = odds ratio
PPV = positive predictive value