


# Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records

Journal of Diabetes Science and Technology  
2016, Vol. 10(1) 6–18  
© 2015 Diabetes Technology Society  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1932296815620200  
dst.sagepub.com  


Jeffrey P. Anderson, ScD, MPH<sup>1</sup>, Jignesh R. Parikh, PhD<sup>1</sup>,  
Daniel K. Shenfeld, PhD, MSc<sup>1</sup>, Vladimir Ivanov, PhD<sup>1</sup>,  
Casey Marks, PhD<sup>1</sup>, Bruce W. Church, PhD<sup>1</sup>,  
Jason M. Laramie, PhD, MS<sup>1</sup>, Jack Mardekian, PhD<sup>2</sup>,  
Beth Anne Piper, MD<sup>2</sup>, Richard J. Willke, PhD<sup>2</sup>,  
and Dale A. Rublee, PhD<sup>2</sup>

## Abstract

**Background:** Application of novel machine learning approaches to electronic health record (EHR) data could provide valuable insights into disease processes. We utilized this approach to build predictive models for progression to prediabetes and type 2 diabetes (T2D).

**Methods:** Using a novel analytical platform (Reverse Engineering and Forward Simulation [REFS]), we built prediction model ensembles for progression to prediabetes or T2D from an aggregated EHR data sample. REFS relies on a Bayesian scoring algorithm to explore a wide model space, and outputs a distribution of risk estimates from an ensemble of prediction models. We retrospectively followed 24 331 adults for transitions to prediabetes or T2D, 2007-2012. Accuracy of prediction models was assessed using an area under the curve (AUC) statistic, and validated in an independent data set.

**Results:** Our primary ensemble of models accurately predicted progression to T2D (AUC = 0.76), and was validated out of sample (AUC = 0.78). Models of progression to T2D consisted primarily of established risk factors (blood glucose, blood pressure, triglycerides, hypertension, lipid disorders, socioeconomic factors), whereas models of progression to prediabetes included novel factors (high-density lipoprotein, alanine aminotransferase, C-reactive protein, body temperature; AUC = 0.70).

**Conclusions:** We constructed accurate prediction models from EHR data using a hypothesis-free machine learning approach. Identification of established risk factors for T2D serves as proof of concept for this analytical approach, while novel factors selected by REFS represent emerging areas of T2D research. This methodology has potentially valuable downstream applications to personalized medicine and clinical research.

## Keywords

electronic health records, diabetes mellitus, type 2, disease progression, medical informatics, prediabetic state

Over 20 million US adults have type 2 diabetes (T2D); prevalence has more than tripled since 1990.<sup>1,2</sup> Prediabetes, an asymptomatic state in which blood glucose concentrations are elevated but lower than diagnostic thresholds, confers high risk for development of T2D. Previous studies have reported demographics, comorbidities, clinical measures, family history, lifestyle, and anthropomorphic measures may be associated with progression.<sup>3</sup> Further elucidation of the factors that drive progression to prediabetes/diabetes would

be valuable in characterizing and intervening on at-risk patients. Prevention and clinical management of patients on

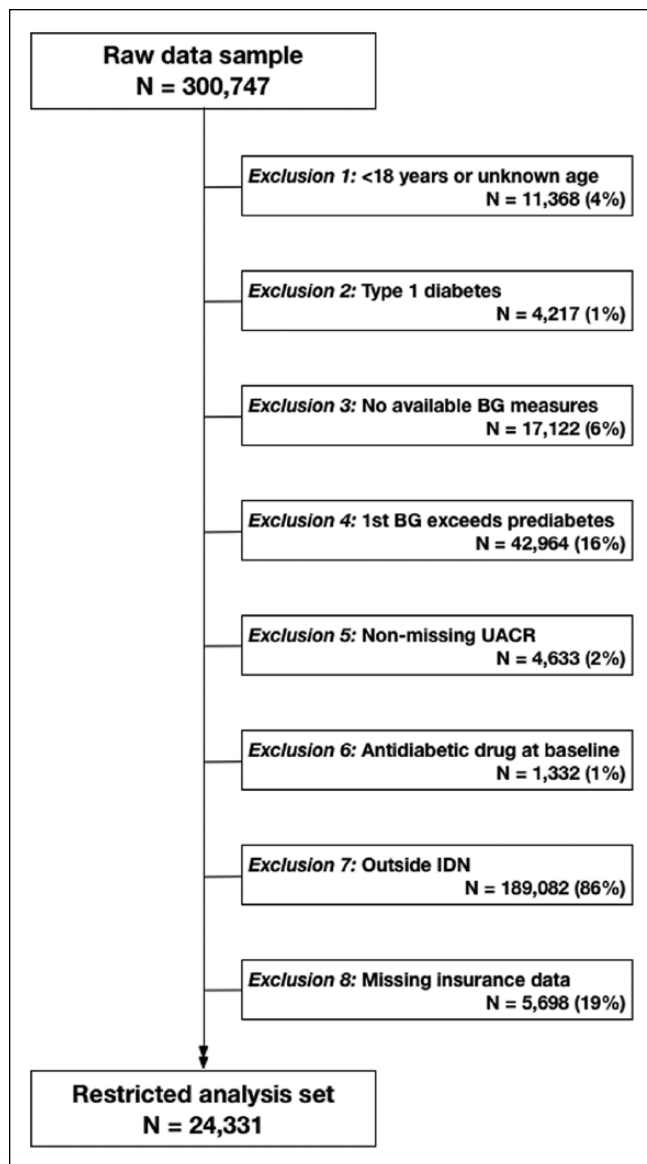
<sup>1</sup>GNS Healthcare, Cambridge, MA, USA

<sup>2</sup>Pfizer Inc, New York, NY, USA

### Corresponding Author:

Jeffrey P. Anderson, ScD, MPH, GNS Healthcare, 196 Broadway,  
Cambridge, MA 02139, USA.

Emails: janderson@gnshealthcare.com and jpa696@mail.harvard.edu



**Figure 1.** Flow diagram describing restriction criteria for analytical study population applied to Humedica electronic health records data sample, 2007-2012. BG, blood glucose; IDN, integrated delivery network; UACR, urinary albumin to creatinine ratio.

the diabetes spectrum could have a major impact on personal and population health, and health care resource utilization and expenditures.

The digitalization of clinical records has provided a rich source of high-dimensional data, and presents a unique opportunity for powerful machine learning approaches to identify patterns and predict outcomes. Several diabetes-related prediction models have been reported, though inconsistency remains—model-building approaches, effect estimates, and the overall accuracy and validation of these prediction models vary to the point that consensus has not been reached.<sup>4-6</sup> Whereas prediction

models are generally constructed in specialized cohorts, with variable selection based on prior publications and/or investigator preconceptions, a hypothesis-free machine learning approach could lead to novel insights into clinical progression and care.<sup>7,8</sup> Specifically, accurate predictions of progression to T2D, based on real-world data, would have distinct value for caregivers and patients with modifiable risk factors. Our objective was to identify patient characteristics that predict progression to prediabetes and T2D in a US adult population, as a practical application of pairing machine learning with electronic health records (EHRs), to characterize disease progression and identify opportunities for intervention.

## Methods

### Data

The source data consisted of clinical records for US adults, 2007-2012, provided by Humedica, Inc (www.humedica.com). Available data included demographic information, ICD-9 codes, prescriptions, laboratory values, and vital signs. Data were deidentified and determined by an independent institutional review board as nonhuman subjects research. Patients eligible for the study population (N = 24 331) were adults belonging to any integrated health care delivery network (IDN), without type 1 diabetes, entering the database with blood glucose measures in the low-risk (normoglycemic) range (Figure 1). The Humedica database includes over 20 IDNs, distributed across all regions of the United States. The exclusion of non-IDN patients (consisting of patients receiving care from various multi-specialty, ambulatory service providers), though substantial, was done to ensure that the record of health care interactions for each study patient would be as complete as possible. Data for non-IDN patients were set aside and used as a testing set for prediction models constructed within the study population.

### Modeling Approach

We evaluated 3 progression models: (1) normoglycemia to T2D, (2) normoglycemia to prediabetes, and (3) prediabetes to T2D. An incident T2D event was defined as the first observed ICD-9 diagnosis code specific to T2D. Patients were considered to have transitioned to prediabetes on the earliest date corresponding to multiple ( $\geq 2$ ) consecutive glucose measures qualifying as prediabetes, according to American Diabetes Association and World Health Organization criteria.<sup>9,10</sup> Specifically, the prediabetes threshold was defined as fasting glucose  $\geq 110$  mg/dL, 2-hour oral glucose tolerance  $\geq 140$  mg/dL, random glucose  $\geq 140$  mg/dL, or hemoglobin A1c (HbA1c)  $\geq 5.7\%$ . Models for progression from prediabetes to T2D were restricted to patients who transitioned to prediabetes as described.

## Variables

The outcomes of interest were prediabetes or T2D. Covariates included demographics (age, gender, race, region, insurance status, 3-digit zip-code-based average annual household income and education level); laboratory values (hemoglobin A1c, fasting glucose, 2-hour oral glucose tolerance, random glucose, triglycerides, total bilirubin, alanine aminotransferase [ALT], creatinine, low-density lipoprotein [LDL], high-density lipoprotein [HDL], C-reactive protein [CRP]); clinical observations (heart rate, blood pressure, body temperature, body mass index [BMI]); ICD-9 diagnosis codes (grouped using Agency for Healthcare Research and Quality Clinical Classifications Software);<sup>11</sup> and prescriptions (National Drug Codes, classified using the Wolters Kluwer Medi-Span Generic Product Identifier groupings; www.medispans.com). The analysis data set consisted of 442 variables. Patients with missing values were not removed; instead variables were modeled as discrete with a missing category, where applicable. This approach was chosen due to its comparability to imputation methods,<sup>12</sup> the ability to retain data while considering many variables (vs a complete-case analysis), and for practical application in clinical assessments where information on key factors may be unknown.

## REFS Bayesian Analytics Platform

We applied a novel analytic platform, Reverse Engineering and Forward Simulation (REFS<sup>TM</sup>) to generate prediction models for progression to diabetes. REFS uses Bayesian inference to learn models directly from data, without pre-specified hypotheses. Instead of a single model, REFS produces an *ensemble* of models sampled from the Bayesian posterior. The ensemble approach has several advantages—applied rigorously, it significantly reduces the risk of overfitting, and provides a framework to estimate distributions of individual variable effects. As chosen from a prior sensitivity analysis, ensembles for this study consisted of 248 individual component models.

To produce each model in the ensemble, REFS scores the posterior probability of a vast number of putative models, using a maximum entropy structural prior as previously described.<sup>13</sup> A model's Bayesian score is approximated by marginalizing out the model parameters and applying the Bayesian Information Criterion, which penalizes complexity. Since the space of possible models is too large to enumerate, REFS uses a Markov Chain Monte Carlo approach to generate samples from the equilibrium distribution of models weighted by their score. Each subsequent evaluation corresponds to a small local transformation, such as adding or removing a single model term. To accelerate convergence, a simulated annealing approach was used to obtain samples from the desired posterior distribution. For further detail on REFS, we refer the reader to the appendix.

Prediction model ensembles,  $\beta$  estimates, predicted probabilities, and area under the (receiver operating characteristic) curve (AUC) estimates were generated using REFS. Supplemental analyses, including Kaplan–Meier plots and multivariable Cox regression models, were conducted using R (version 2.15.0). Cox model estimates are reported as hazard ratios (HRs) with 95% confidence intervals (CIs). Effect estimates were adjusted for factors selected by REFS, in addition to available diabetes risk factors identified a priori.<sup>3,4</sup>

## Results

We evaluated 24 331 eligible patients for progression outcomes. During follow-up, 15% (N = 3765) were diagnosed with T2D. Transition to prediabetes was observed in 46% of the study population. The rate of progression from normoglycemia to T2D was 4.72 events per 100 person-years; normoglycemia to prediabetes, 18.72 events per 100 person-years; prediabetes to T2D, 8.6 events per 100 person-years. Distributions of baseline characteristics by T2D status are listed in Table 1. Patients in the study population were likely to be female, Caucasian, and from the Midwest region. The distributions of T2D events suggested positive associations with: male gender, older age, African American race, South region, inconsistent insurance coverage, low income, hypertension, obesity, higher blood glucose at baseline, high triglycerides, and dyslipidemia (Table 1).

Consistent predictors for the 3 progression model ensembles are summarized in Table 2. We first evaluated predictors in each REFS ensemble by identifying the proportion of models that include each factor (selection frequency). For progression from normoglycemia to T2D, factors that were selected in every component model included blood glucose (test-specific tertiles), hypertension, income, insurance status, race, and triglycerides. Additional factors frequently selected were lipid disorders (97%), and systolic blood pressure (77%). Those with high baseline blood glucose progressed to T2D nearly 3 times faster on average, relative to those in the lowest category (HR = 2.95, 95% CI: 2.69, 3.23; Table 2, Figure 2). We observed a dose-response relationship between triglycerides and progression to T2D (Table 2, Figure 2). Patients with hypertension (HR = 1.33, 95% CI: 1.23, 1.44) or lipid disorders (HR = 1.18, 95% CI: 1.08, 1.29) progressed faster, and self-reported race (African American vs Caucasian, HR = 1.60, 95% CI: 1.47, 1.75) predicted progression to T2D (Table 2).

To evaluate predictive performance of the ensemble, we calculated AUC statistics. For the ensemble predicting progression from normoglycemia to T2D, AUC was 0.76, reflecting moderately strong accuracy in predicting T2D (Figure 3). To assess performance of the model outside of the training data, we tested the ensemble in a separate data set (the non-IDN population; N = 189 082). In this testing set, AUC of the ensemble predicting progression to T2D from

**Table 1.** Distributions of Selected Baseline Characteristics Among the Primary Study Population (N = 24 331), and Proportion Progressing to Type 2 Diabetes, Humedica Electronic Health Records Data Sample, 2007-2012.

Variable	n (%)	Diabetes events, n (%) <sup>a</sup>
Gender		
Female	15 272 (63)	2220 (15)
Male	9059 (37)	1545 (17)
Age (years)		
≤30	2246 (9)	181 (8)
31-45	4842 (20)	559 (12)
46-60	7234 (30)	1305 (18)
61-70	3659 (15)	805 (22)
>70	6350 (26)	915 (14)
Race		
Caucasian	14 836 (61)	2181 (15)
African American	3557 (15)	900 (25)
Asian	505 (2)	74 (15)
Other/unknown	5433 (22)	610 (11)
Region		
Midwest	22 362 (92)	3230 (14)
South	1858 (8)	523 (28)
West/Northeast	111 (<1)	12 (11)
Insurance status		
Commercial	11 317 (47)	1452 (13)
Medicare	6822 (28)	779 (11)
Inconsistent	5866 (24)	1491 (25)
Uninsured/other	326 (1)	43 (13)
Average income (US\$/year)		
<40 000	6021 (25)	1371 (23)
40 000-55 000	13 877 (57)	1607 (12)
>55 000	4433 (18)	787 (18)
Systolic blood pressure (mmHg)		
<90	87 (<1)	7 (8)
90-119	4077 (17)	594 (15)
120-139	4645 (19)	1045 (22)
≥140	2378 (10)	665 (28)
Missing	13 144 (54)	1454 (11)
Body mass index (kg/m <sup>2</sup> )		
<18.5	93 (<1)	12 (13)
18.5-24.9	1849 (8)	173 (9)
25.0-29.9	2581 (11)	405 (16)
30.0-34.9	1719 (7)	405 (24)
≥35	1562 (6)	445 (28)
Missing	16 527 (68)	2325 (10)
Family history of diabetes		
No evidence	24 290 (>99)	3753 (15)
Yes	41 (<1)	12 (29)
Hypertension		
No evidence	20 344 (84)	2543 (13)
Yes	3897 (16)	1222 (31)
Lipid disorders		
No evidence	21 013 (86)	2852 (14)
Yes	3318 (14)	252 (28)

(continued)

**Table 1. (continued)**

Variable	n (%)	Diabetes events, n (%) <sup>a</sup>
Blood glucose		
Low	8453 (35)	914 (11)
Medium	9647 (40)	1616 (17)
High	6231 (26)	1235 (20)
Triglycerides (mg/dL)		
≤150	11 796 (48)	1749 (15)
151-199	2341 (10)	542 (23)
200-499	2257 (9)	589 (26)
≥500	127 (1)	34 (27)
Missing	7810 (32)	851 (11)
High-density lipoprotein (mg/dL)		
<50 (female)/<40 (male)	6108 (25)	1193 (20)
50-59 (female)/40-59 (male)	5653 (23)	1010 (18)
≥60	4573 (19)	707 (15)
Missing	7997 (33)	855 (11)

<sup>a</sup>Percentages are row percentages, that is, the proportion of category-specific patients with a T2D event.

normoglycemia was 0.78, indicating consistency with the training data. For context, investigators of the Framingham Offspring Study reported an AUC of 0.72 for their “personal model” (consisting of variables that would generally be known to a patient, ie, age, sex, parental history of diabetes, BMI), and an AUC of 0.85 when clinical variables (oral glucose tolerance, fasting insulin, CRP, and indexes indicating insulin sensitivity/resistance) were included.<sup>6</sup> Our ensemble of prediction models exhibited comparable performance, using EHR data, even though some key predictors used by the Framingham investigators were not uniformly available (ie, family history of diabetes, waist circumference).

In addition, we generated individual diabetes risk distributions for 2 patients with contrasting covariate profiles (Figure 4). Patient 59 was a 48-year-old Caucasian female from a high-income area (mean = \$68 000/year), with a baseline random glucose of 81 mg/dL, low triglycerides (73 mg/dL), without hypertension or a lipid disorder. Her corresponding 3.5-year risk of progressing to T2D ranged from 6% to 10%, with a mean of 9.2%. Patient 5076 was a 46-year-old African American female from a low-income area (mean = \$32 000/year), with a baseline random glucose of 128 mg/dL, high triglycerides (508 mg/dL), hypertension, and a lipid disorder. Her predicted 3.5-year risk of progressing to T2D ranged from 70% to 84% (mean = 77.0%). During follow-up, patient 5076 progressed to T2D, while patient 59 did not.

A summary of the predictors of progression to prediabetes can be found in Table 2. Consistent with the T2D model, baseline blood glucose and insurance status were selected in every component model. Additional predictors of prediabetes included age (100%), body temperature (100%), ALT

**Table 2.** Variable Selection Frequency and Effect Estimates for Selected Patient Factors Across 3 Models of Progression to Prediabetes or T2D, Humedica Electronic Health Records Data Sample, 2007-2012.

Variable <sup>a</sup>	Normal → T2D			Normal → prediabetes			Prediabetes → T2D		
	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)
Age (years)									
≤30	—	—	—	100	Referent	Referent	—	—	—
31-45					0.108	1.41 (1.27, 1.57)			
46-60					0.221	2.18 (1.96, 2.41)			
61-70					0.248	2.61 (2.34, 2.92)			
>70					0.260	2.82 (2.52, 3.16)			
Race	100	Referent	Referent	—	—	—	31	Referent	Referent
Caucasian									
African American		0.081	1.60 (1.47, 1.75)					0.020	1.31 (1.14, 1.50)
Asian		0.029	1.30 (1.03, 1.65)					0.015	1.28 (0.92, 1.79)
Other		-0.009	0.99 (0.90, 1.08)					-0.007	0.87 (0.76, 1.00)
Insurance	100	Referent	Referent	100	Referent	Referent	100	Referent	Referent
Commercial									
Medicare		-0.003	0.87 (0.78, 0.97)						
Inconsistent		0.077	1.31 (1.20, 1.42)						
Other		0.035	1.57 (1.15, 2.14)						
Average income (US\$1000/yr)	100	Referent	Referent	—	—	—	100	Referent	Referent
<40									
40-54.9		-0.055	0.64 (0.59, 0.69)						
≥55		-0.009	0.92 (0.84, 1.01)						
Hypertension	100	0.064	1.33 (1.23, 1.44)	—	—	—	100	0.079	1.51 (1.34, 1.69)
Lipid disorders	97	0.042	1.18 (1.08, 1.29)	—	—	—	—	—	—
Body mass index (kg/m <sup>2</sup> )	—	—	—	81	0.043	1.25 (0.90, 1.73)	—	—	—
<18.5									
18.5-24.9					Referent	Referent			
25.0-29.9					0.052	1.16 (1.04, 1.29)			
30.0-34.9					0.083	1.33 (1.19, 1.49)			
≥35					0.129	1.70 (1.51, 1.91)			

(continued)

**Table 2. (continued)**

Variable <sup>a</sup>	Normal → T2D			Normal → prediabetes			Prediabetes → T2D		
	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)	Selection frequency (%) <sup>b</sup>	Average beta estimate <sup>c</sup>	Adjusted <sup>d</sup> HR (95% CI)
Body temperature >98.6°F	—	—	—	100	0.114	1.40 (1.24, 1.58)	—	—	—
Blood glucose	100	—	—	100	—	—	—	—	—
Low	—	Referent	Referent	—	Referent	Referent	—	—	—
Medium	—	0.052	1.52 (1.40, 1.65)	—	0.060	1.26 (1.20, 1.32)	—	—	—
High	—	0.146	2.95 (2.69, 3.23)	—	0.136	1.86 (1.76, 1.96)	—	—	—
Triglycerides (mg/dL)	100	—	—	19	—	—	98	—	—
≤150	—	Referent	Referent	—	Referent	Referent	—	Referent	Referent
151-199	—	0.070	1.40 (1.27, 1.54)	—	0.005	1.04 (0.97, 1.11)	—	0.063	1.30 (1.12, 1.50)
200-499	—	0.082	1.44 (1.30, 1.59)	—	0.010	1.12 (1.04, 1.19)	—	0.080	1.31 (1.13, 1.51)
≥500	—	0.118	2.02 (1.42, 2.88)	99	0.020	1.21 (0.93, 1.57)	—	0.136	2.24 (1.36, 3.68)
Elevated ALT <sup>e</sup>	—	—	—	78	0.046	1.19 (1.13, 1.25)	—	—	—
C-reactive protein (mg/dL)	—	—	—	—	Referent	Referent	—	—	—
<1	—	—	—	—	0.027	1.21 (1.01, 1.44)	—	—	—
1-3	—	—	—	—	0.086	1.64 (1.40, 1.91)	—	—	—
>3	—	—	—	—	—	—	—	—	—
HDL <sup>f</sup>	—	—	—	19	—	—	—	—	—
Low	—	—	—	—	Referent	Referent	—	—	—
Medium	—	—	—	—	-0.003	0.86 (0.81, 0.91)	—	—	—
High	—	—	—	—	-0.008	0.76 (0.71, 0.81)	—	—	—

Abbreviations: ALT, alanine aminotransferase; CI, confidence interval; HDL, high-density lipoprotein; HR, hazard ratio; T2D, type 2 diabetes.

<sup>a</sup>Percentage missing: age, 0%; race, 0%; insurance, 0%; income, 0%; hypertension, 0%; lipid disorders, 0%; BMI, 68%; body temperature, 89%; blood glucose, 0%; triglycerides, 32%; ALT, 20%; CRP, 94%; HDL, 33%.

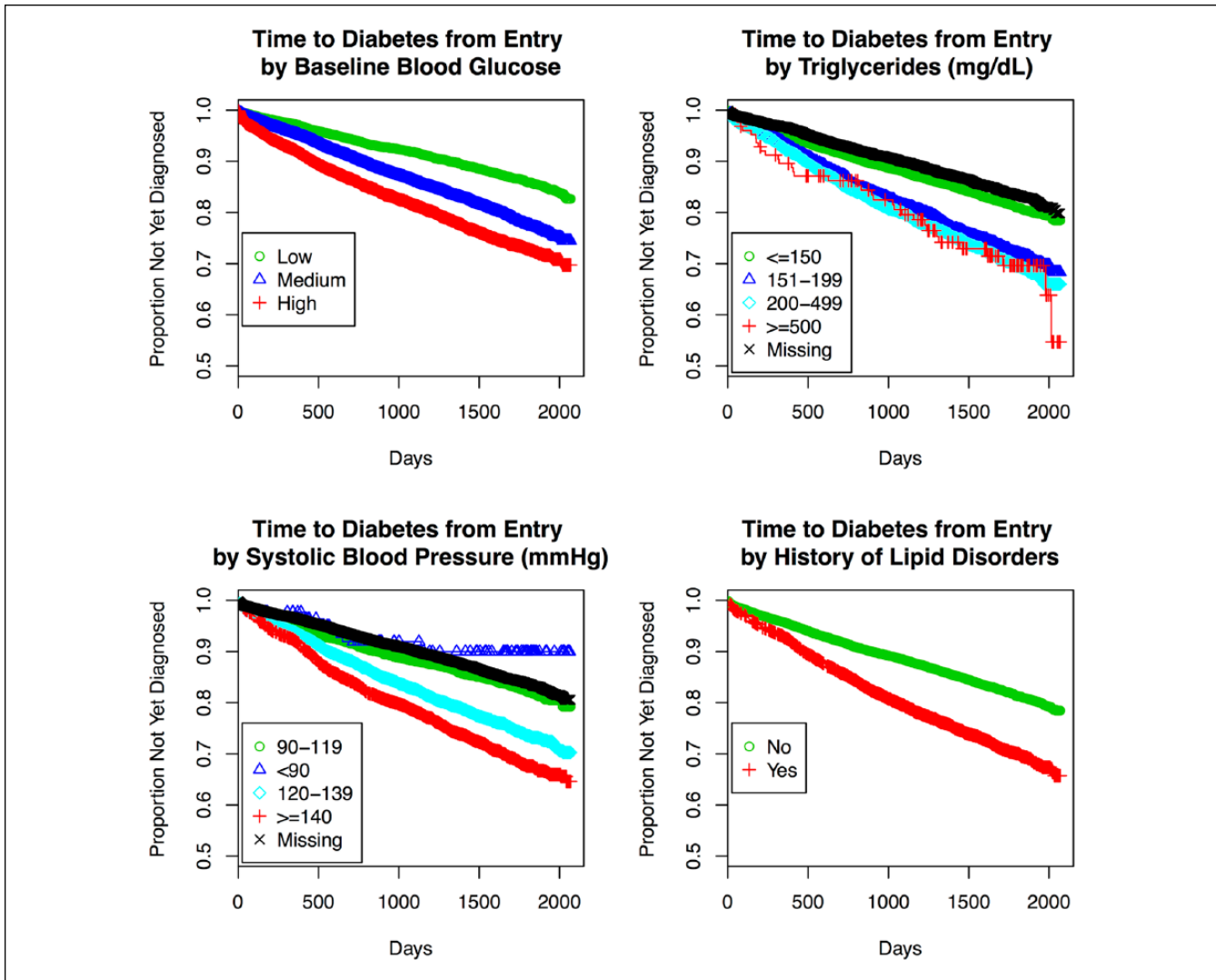
<sup>b</sup>Percentage of individual models in the REFS ensemble that included the specified variable.

<sup>c</sup>Mean change in log odds of (pre)diabetes (vs referent) across the ensemble.

<sup>d</sup>Multivariable Cox regression model estimates are adjusted for the following listed covariates. Progression from normal to T2D: blood glucose, blood glucose assay, hypertension, income, insurance, race, triglycerides, lipid disorders, systolic blood pressure, gender, age, family history of diabetes, LDL, BMI, and antihyperlipidemic therapy. Progression from normal to prediabetes: age, blood glucose, blood glucose assay, body temperature, insurance, ALT, BMI, CRP, HDL, triglycerides, family history of diabetes, hypertension, income, race, and LDL. Progression from prediabetes to T2D: hypertension, income, insurance, triglycerides, heart disease, cerebrovascular disease, race, diastolic blood pressure, gender, age, family history of diabetes, blood glucose, BMI, LDL, and HDL.

<sup>e</sup>For males: ALT was considered to be elevated if ≥50 IU/mL; females: ≥38 IU/mL.

<sup>f</sup>Males: low HDL, <40 mg/dL; high HDL, ≥60 mg/dL. Females: low HDL, <50 mg/dL; high HDL, ≥60 mg/dL.

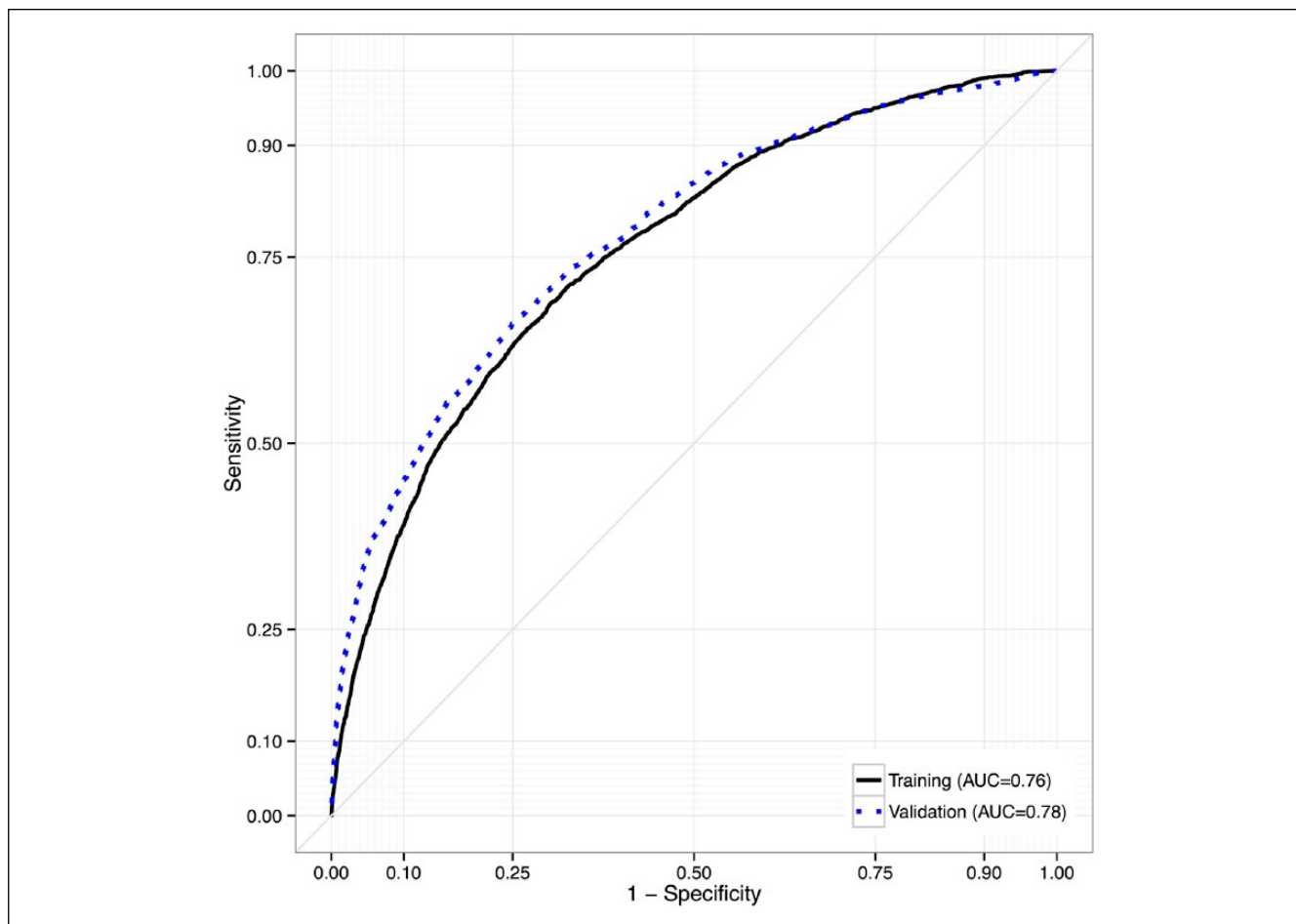


**Figure 2.** Kaplan–Meier plots for time to T2D by selected (potentially modifiable) patient factors: baseline blood glucose measures, triglycerides, systolic blood pressure, and history of lipid disorders, Humedica electronic health records data sample, 2007-2012.

(99%), BMI (81%), CRP (78%), triglycerides (19%), and HDL (19%; Table 2). Specifically, positive dose-dependent associations with prediabetes were identified for age, baseline blood glucose, BMI, and CRP. We also observed a negative association between rate of progression to prediabetes and HDL (Table 2). In addition, higher baseline body temperature, Medicare coverage, and elevated ALT were associated with faster progression to prediabetes (Table 2). Of note, missingness of some variables (ALT, BMI, CRP, HDL, triglycerides) appeared to be associated with progression (Table 2). Whereas many of the selected predictors are recognizable as established risk factors for diabetes, others (ALT, HDL, CRP, and body temperature) may be novel in that they represent plausible but still emerging areas of diabetes research and targets for intervention. Kaplan–Meier plots for time to prediabetes by these factors are displayed in Figure 5. The corresponding AUC for the ensemble

predicting progression from normoglycemia to prediabetes was 0.70 (testing set, AUC = 0.72).

Table 2 also summarizes the prediction model ensemble for progression from prediabetes to T2D ( $n = 10\,616$ ). Hypertension, income, and insurance status were uniformly represented. Additional predictors included triglycerides (98%), heart disease (92%), cerebrovascular disease (73%), race (31%), and diastolic blood pressure (29%). Positive associations with progression from prediabetes to T2D were observed for hypertension, low income, triglycerides, and African American race. Whereas this baseline model ensemble performed reasonably well in predicting progression within the training set (AUC = 0.71), it was not replicated in the testing set (AUC = 0.58). A possible explanation for this phenomenon could be a higher proportion of unspecified information on race in the non-IDN study population (51% vs 24%).



**Figure 3.** Receiver operating characteristic curves for accuracy of the REFS ensemble in predicting progression to diabetes (from normoglycemia) in the training and testing data sets, Humedica electronic health records data sample, 2007-2012.

## Discussion

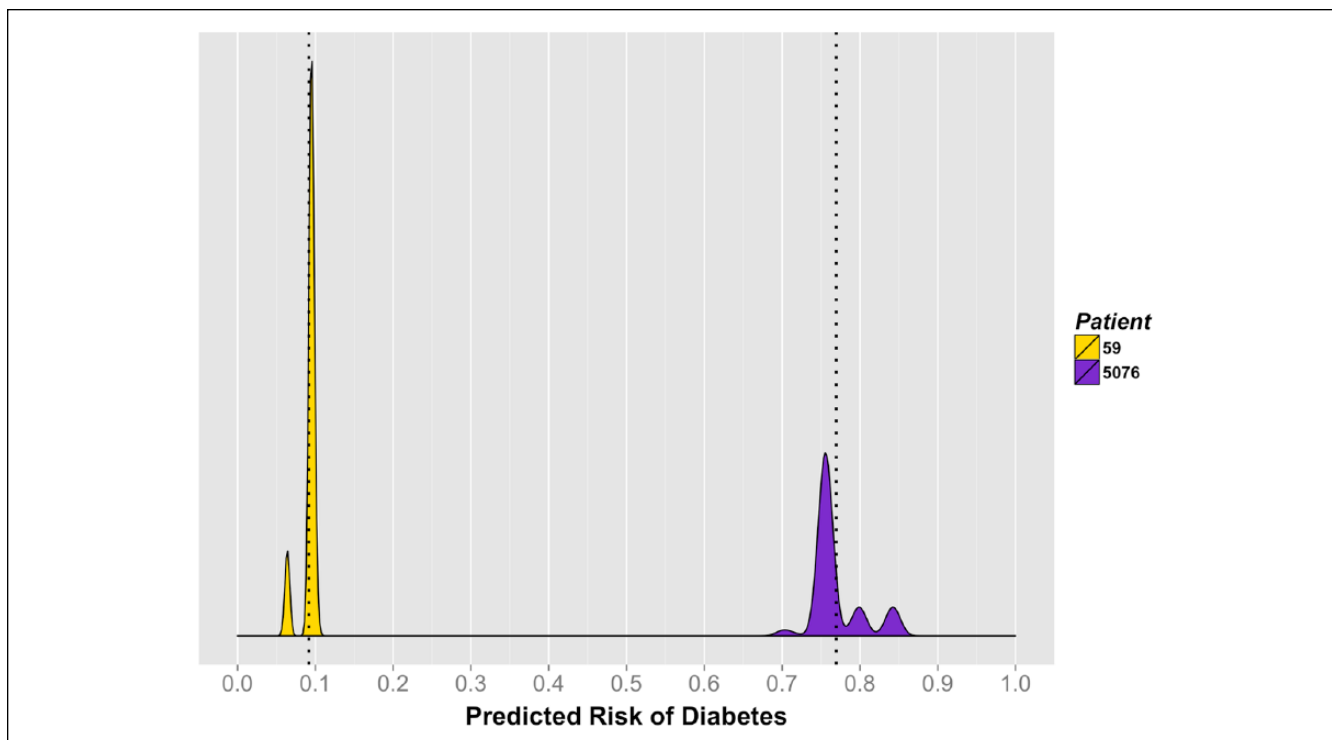
Using a hypothesis-free machine learning ensemble approach, we have constructed a series of prediction models that identify the patient factors most associated with progression to T2D in real-world EHR data. Despite inherent noise that generally afflicts such data, predictive accuracy was relatively strong. The ensemble for progression from normoglycemia to T2D showed high predictive value in particular, and was replicated in the testing data set. Performance of our models was comparable to prediction models that have previously been reported (range, 0.71 to 0.88), though a validation study conducted by Abbasi et al suggested that nearly all models considered had overestimated T2D risk.<sup>5</sup> Given the size of our data sample, we expect that accuracy would be stable across external data sets, though further validation is warranted.

Although the REFS platform learns models directly from data without prespecified hypotheses, several patient factors considered to be established correlates of T2D were selected. Specifically, blood glucose measures, age, race, triglycerides, BMI, and blood pressure/hypertension have consistently been

identified as risk factors for development of T2D, and were confirmed to varying degrees in our study.<sup>3</sup> Identification of such factors serves to qualitatively validate both the analytical methods and the source data, while strengthening the body of evidence that these factors are mechanistically linked to diabetes progression. Conversely, other factors previously thought to associate with T2D were not replicated here. Some of these variables may not have been selected because they were not widely available (ie, family history, lifestyle factors); others may not have additional explanatory value once other factors are accounted for (ie, gender). We expect that more complete data on relevant covariates would further improve the accuracy of similar prediction models.

In addition to established risk factors, relatively novel predictors were also identified, particularly in prediabetes models. First, HDL was consistently selected throughout progressive iterations of modeling. Although HDL only appeared in 19% of primary ensemble models for progression to prediabetes, results from survival analysis suggested a moderate inverse relationship between HDL and rate of progression to prediabetes, including a 24% slower rate of





**Figure 4.** Individual 3.5-year risk of diabetes for 2 selected patients, Humedica electronic health records data sample, 2007-2012.

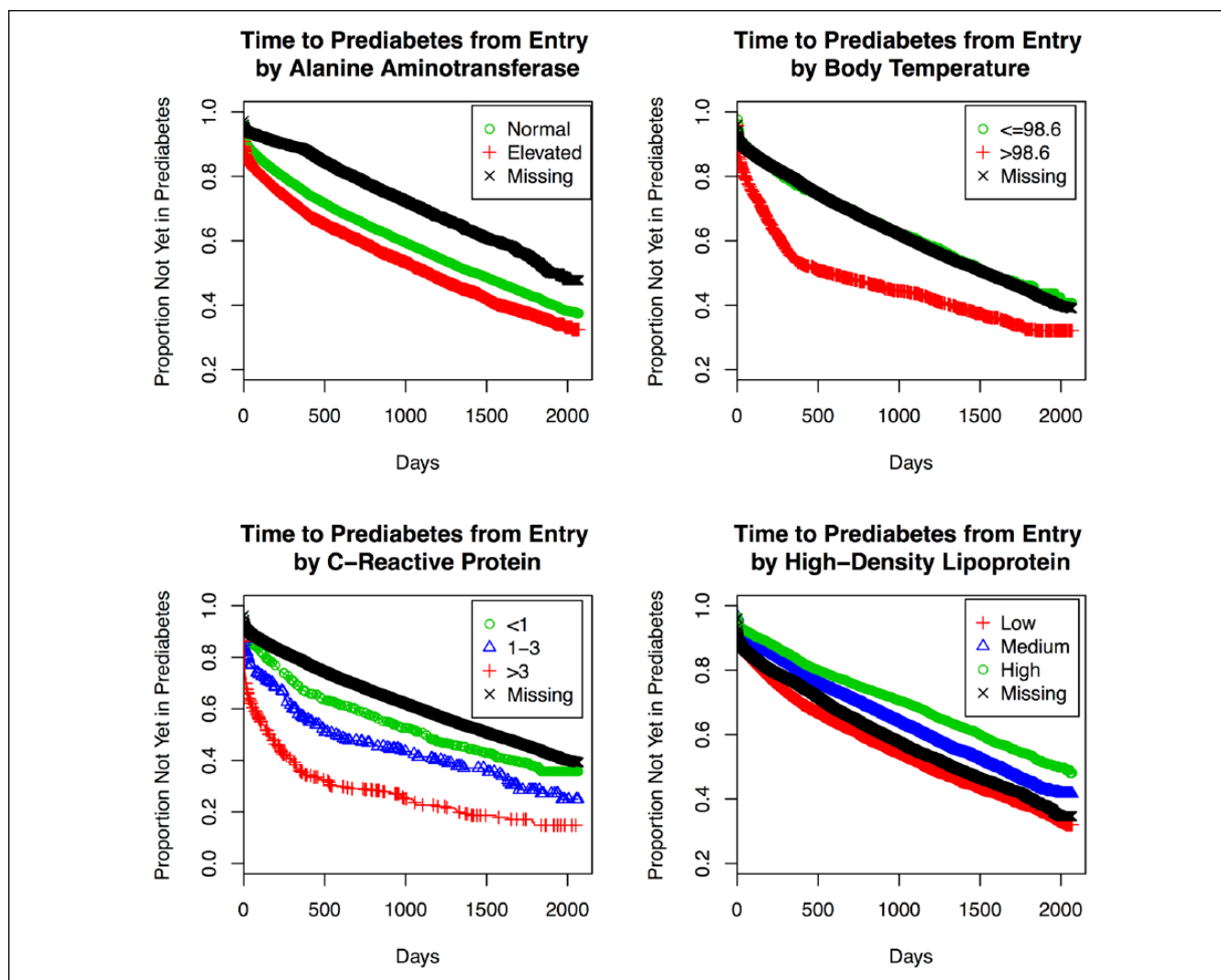
progression among those within the highest level of HDL versus the (gender-specific) lowest HDL category. This finding is consistent with emerging evidence for a role for HDL in diabetes pathophysiology, with several plausible mechanisms of action having been proposed.<sup>14</sup> An association between the B1B1 Taq1B genotype, which leads to marked decrease in HDL levels, and T2D has been reported (odds ratio, 1.83; 95% CI, 1.12, 2.99).<sup>15</sup> In a post hoc analysis of the ILLUMINATE trial, a lower proportion of participants on atorvastatin plus torcetrapib (a drug that elevates HDL levels) developed T2D, relative to those on atorvastatin alone.<sup>16</sup> This difference was only marginally significant ( $P = .09$ ), likely due to few events. Taken together, HDL represents an intriguing opportunity for identification of individual diabetes risk and for clinical intervention.

Elevated ALT was also selected as a predictor of prediabetes, associated with a 19% faster rate of progression. A hepatic enzyme, ALT is used as a biomarker to assess severity of liver dysfunction, and some investigators have suggested a link with T2D. Vozarova et al reported a nearly 2-fold increased hazard of diabetes associated with the 90th versus 10th percentile of ALT (HR = 1.9, 95% CI: 1.1, 3.3).<sup>17</sup> In a recent meta-analysis, investigators calculated a pooled adjusted relative risk of diabetes of 1.26 (95% CI: 1.14, 1.41) per standard deviation change in log-transformed ALT. However, after a statistical correction for publication bias was applied, this estimate became nonsignificant.<sup>18</sup> Thus the role of ALT in diabetes is not yet established. ALT is a marker

of nonalcoholic fatty liver disease, which has been associated with insulin resistance, therefore the connection is plausible.<sup>19</sup> Our findings suggest that the magnitude of the effect of ALT may be low, and may be unique in having a large enough sample size to detect this association.

Two other factors, CRP and body temperature, also emerged as predictors of prediabetes, although these measures were not consistently documented (6% and 11% respectively). CRP, a marker of systemic inflammation, may be elevated in obese individuals, and has been linked to T2D. Investigators in the Rotterdam Study reported a 67% increased hazard of T2D in individuals with elevated CRP, consistent with our findings ( $>3$  vs  $<1$  mg/dL, HR = 1.63, 95% CI: 1.39, 1.90), and estimated that one-third of diabetes in the Dutch population may be attributable to CRP.<sup>20</sup> Evidence for a mechanistic role of body temperature in diabetes is currently limited. A recent study in a rodent model demonstrated that direct injection of insulin into the hypothalamus resulted in dose-dependent increases in core body temperature.<sup>21</sup> As progressive declines in insulin sensitivity lead to greater levels of circulating insulin, a link between diabetes pathophysiology and higher core body temperature is plausible. Further research is warranted to establish the clinical significance of CRP and body temperature in diabetes.

Limitations of our study were primarily related to the availability of data. First, patients were not observed uniformly. Changes in status are not always recorded, and availability of clinical data itself may not be a random process. In some cases,



**Figure 5.** Kaplan–Meier plots for time to prediabetes by selected patient factors, Humedica electronic health records data sample, 2007-2012.

variables that are likely to be associated with diabetes (ie, family history, smoking, adiposity measures) were not widely available in the data sample. For others (ie, BMI, CRP), available measures were not frequently updated. It is likely that a similar analysis in a more comprehensive data set would result in improved accuracy for resulting prediction models. Still, our prediction model ensembles performed strongly in an EHR sample that is representative of real-world clinical data. Last, interpretation of our baseline prediction models may be somewhat limited in that they were not designed to assume direct causality, and patient measures were not updated in follow-up.

Although specific to diabetes, our study illustrates the value of applying machine learning to identify and characterize risk factors for health care outcomes from EHR data. These methods could be especially valuable in contexts where disease processes or interventions are not well established. Last, our method of real-world data-driven modeling could have an

impact on health care by identifying at-risk patients early enough to present opportunities for prevention and clinical management. For example, in this report we closely examined 2 patients, and projected their individual risk profiles. Incorporation of personalized risk profiles such as these into standard clinical evaluation could have potential for increasing the specificity and success of targeted interventions.

## Conclusions

We constructed accurate prediction model ensembles for progression to T2D using a novel machine learning platform based on Bayesian mathematics and an extensive EHR database. These results confirmed established risk factors for T2D and identified novel factors, for which roles in diabetes pathophysiology are plausible. Our approach has potential wide-ranging applications in several disease areas, and could

be developed into powerful tools for health care research, as well as for clinical applications related to personalized risk assessment and targeted interventions.

## Appendix

### Building Predictive Model Ensembles Using REFS

**Model Space.** Modern “big-data” problems suffer from “the curse of dimensionality”: the number of variables is very large, and the dimension of the space of statistical models depends on it exponentially. This means that any statistical model is necessarily underdetermined, even if we have many observations. In this situation, even the best model has relatively little explanatory power. A human expert could perhaps handle several dozen models, but this is clearly not a scalable solution; and generating random models using predetermined hypotheses generally leads to bias and overfitting. Instead, REFS aims to discover the data generating process by building models in a hypothesis-free way.

We start by defining the hypothesis space under consideration. This should be a rich enough space to plausibly contain models that are close to the “true” data generating process, but not so large as to become computationally intractable. The reverse engineering (RE) part of the REFS procedure explores this space, and the result is an ensemble of Bayesian predictive models. In the second stage, models are queried using forward simulation (FS), to produce predictions and study inference questions.

Specifically, consider a multivariate system  $X = (X_1, \dots, X_N)$ , where each random variable may take on values from a discrete or continuous domain, and an outcome variable  $y$ . As our hypothesis space we choose a space of generalized linear models,<sup>22</sup> where each model  $\mathbf{M} = (M, \Theta_M)$  consists of an interaction form  $M$  and parameters  $\Theta$ . The interaction form is a polynomial function in the covariates  $X_i$ , and it determines which covariates and which interactions between them are part of the model. In our current implementation we allow cross terms, but not higher order interactions. For a given interaction form  $M$  we denote by  $l(M)$  the total number of terms in  $M$  (linear and interactions), and by  $L$  the total number of possible terms and cross terms derived from  $X$ .

The interaction form  $M$  determines the design matrix for the generalized linear model  $\mathbf{M}$ . The link function is chosen based on the type of the outcome; for binary outcomes for this study we use a logit link function. The parameters of the model are encoded by  $\Theta$ .

**Model Scoring.** Under the Bayesian framework, we are interested in the Bayesian posterior of a model  $\mathbf{M}$  given data  $D$ :

$$P(\mathbf{M}|D) \propto P(D|\mathbf{M})P(\mathbf{M}) = \sum_M P(D|M)P(M) \quad (1)$$

$$P(D|M) = \int P(D|M, \Theta_M)P(\Theta_M|M)d\Theta_M \quad (2)$$

In (1),  $P(M)$  is the *structural prior*, encoding our assumptions about the structure of the interaction form (number of terms, interaction terms, etc) before seeing the data. We use the following *maximum entropy* prior with respect to the average number of terms in  $M$ ,<sup>13,23</sup> which also acts as a strong regularizer:

$$P(M) \propto \left( \frac{L}{l(M)} \right)^{-1}$$

In (2),  $P(\Theta_M)$  is the *parameter prior*, encoding our assumptions about the parameters of the model given the interaction form. We use noninformative parameter priors when the integral (2) can be solved analytically.<sup>24</sup> When a closed-form solution is not available, we may approximate the Bayesian integral via Schwartz’s Bayesian information criterion (BIC).<sup>25</sup>

$$-\log P(D|M) \approx S_{MLE}(M) + \frac{\kappa(M)}{2} \log n$$

where  $S_{MLE}$  is the maximum log-likelihood, computed via an iteratively reweighted least squares (IRLS) method;  $\kappa(M)$  is the number of model parameters, also equal to the number of columns in the design matrix (and closely related, though not necessarily equal to,  $l(M)$ );  $n$  is the number of observations.

A more accurate approximation can be obtained following Gull<sup>26</sup> as follows. Once the MLE for  $\Theta$  has been found, we use a Laplace approximation for the likelihood function. We choose a maximum entropy prior for  $\Theta$ , while using the mean of the outcome  $y$  as a constraint. The result is a Gaussian prior, and the integral against the approximated likelihood function is analytic.

**Model Sampling.** Since the space of models is combinatorially large, the sum (1) cannot be evaluated in practice. To approximate it, we use a Monte Carlo method and build an ensemble of models representing a sample from the Bayesian posterior.<sup>27-29</sup> Strong signal in the data leads to low model uncertainty, since few models contribute meaningfully to the posterior. The variance in predictions obtained from the ensemble will therefore be low. Conversely, a weak signal leads to much broader posterior distributions. Thus, the ensemble method naturally provides a measure of the uncertainty in the predictions we make. The details on the construction of the ensemble are explained in the next section.

Each member of the ensemble consists of the interaction form  $M$  alone; the parameters  $\Theta_M$  in (2) are integrated out as explained above. To produce the ensemble, the space of models is sampled using a Markov chain Monte Carlo (MCMC) method to generate samples from an equilibrium Boltzmann distribution  $\pi$  of candidate model structures from  $P(M|D)$ .<sup>13,29</sup> Each step in the Metropolis Markov chain corresponds to a small perturbation such as adding or deleting a term from the model. Let  $q(M \rightarrow M')$  be the probability of proposing the transition from a model  $M$  to  $M'$ . We

define the acceptance probability according to the Metropolis criterion,<sup>30</sup>

$$P(\text{Accept}(M \rightarrow M')) = \min\left(1, \frac{\pi(M')q(M' \rightarrow M)}{\pi(M)q(M \rightarrow M')}\right)$$

and the transition probability by  $p(M \rightarrow M') = q(M \rightarrow M')P(\text{Accept}(M \rightarrow M'))$ . It is now easy to check the *detailed balance* condition

$$\pi(M)p(M \rightarrow M') = \pi(M')p(M' \rightarrow M),$$

guaranteeing that  $\pi$  is indeed the stationary distribution for the Markov chain.<sup>31</sup>

To accelerate convergence we use a simulated annealing procedure, where we apply the Metropolis method to a sequence of Boltzmann distributions with

$$Pr(M) = Pr(M | D, T_j) \propto \exp(-S(M)/T_j),$$

and decreasing annealing temperature  $T_j$ . Here,

$$S(M) = -\log P(D | M) - \log P(M),$$

where the first term is approximated as described in the previous section and  $P(M)$  is the structural prior. At each stage  $j$  the equilibrated samples from  $T_j$  initialize the Metropolis method at  $T_{j+1}$ . The initial temperature is chosen sufficiently high that probability distribution over models is flat and there are no score barriers to the acceptance, whereas the last temperature is  $T = 1$ , where we are in fact sampling from the Bayesian posterior.

The cooling schedule for the simulated annealing is determined self-consistently from  $P(M|D, T)$  by maintaining a fixed overlap between  $P(M|D, T')$  and  $P(M|D, T)$  where  $T$  and  $T'$  are the current and next temperatures in the cooling schedule. When  $P(M|D, T)$  is changing rapidly with  $T$ , this cooling scheduling takes small steps concentrating the sampling where the problem is most difficult. More precisely, we select  $T'$  so that the overlap, defined as

$$\Omega(T, T') = \exp\left(-\left(S - S_{\min}\right)\left(T'^{-1} - T^{-1}\right)\right),$$

is equal to a predetermined value, 80% in our implementation (note that  $\Omega$  is a monotonic function of  $0 < T' \leq T$ , with values in the interval  $[0, 1]$ ). This process of maintaining overlap helps ensure that the sampling will be correct when  $T = 1$  is reached.

**Inference.** Once an ensemble of models has been produced, samples from the posterior distribution may be obtained by selecting a random model structure  $M$  and values for  $\Theta_M$  from the parameter posterior. Drawing a large number of samples, we readily derive estimators for various properties

of the posterior. Namely, to estimate a property  $f$  of the posterior, dependent on  $\mathbf{M} = (M, \Theta)$ , we have

$$P(f | D) \approx \frac{\sum_{\mathbf{M}} P(\mathbf{M}|D) f(\mathbf{M})}{\sum_{\mathbf{M}} P(\mathbf{M}|D)}$$

where the sum is over the posterior draws.<sup>13,27-29</sup> A simple example is the recovery of training data: setting the values of input variables in the model to values in the training data set, and comparing them to the experimentally observed values. By holding out a portion of the data as a test set, we can obtain out-of-sample predictions and compare them to in-sample results to ensure that no overfitting occurs.

A more interesting application for the analysis of the posterior is given by the interventional derivative. For a continuous covariate  $X$ , we can obtain samples from the posterior of  $y$  conditioned on  $X$  by drawing samples while holding  $X$  constant. The response derivative is then defined as

$$\beta_x = \frac{\partial P(y | X = x)}{\partial x}$$

For a categorical variable  $X$  taking values in  $\{x_0, x_1, \dots, x_m\}$  we can instead compute the difference derivative

$$\beta_{ij} = P(y | X = x_i) - P(y | X = x_j)$$

These statistics and their distributions are a measure of the effect of a given covariate on the outcome in the full ensemble.

## Abbreviations

ALT, Alanine aminotransferase; AUC, area under the curve; BMI, body mass index; CI, confidence interval; CRP, C-reactive protein; EHR, electronic health record; HbA1c, hemoglobin A1c; HDL, high-density lipoprotein; HR, hazard ratio; IDN, integrated delivery network; LDL, low-density lipoprotein; REFS, Reverse Engineering and Forward Simulation; T2D, type 2 diabetes.

## Acknowledgments

The authors would like to thank Ngoc Thai and Karl Runge for technical assistance.

## Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: All authors were employed by either GNS Healthcare or Pfizer at the time this research was conducted.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described in this report was funded by Pfizer, Inc. No financial support was received from any public or not-for-profit agency.

## References

1. Inzucchi SE. Clinical practice: diagnosis of diabetes. *N Engl J Med*. 2012;367:542-550.
2. Gregg EW, Li Y, Wang J, et al. Changes in diabetes-related complications in the United States, 1990-2010. *N Engl J Med*. 2014;370:1514-1523.
3. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. *Lancet*. 2012;379:2279-2290.
4. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9:103.
5. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ*. 2012;345:e5900.
6. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med*. 2007;167(10):1068-1074.
7. Xing H, McDonagh PD, Bienkowska J, et al. Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Comput Biol*. 2011;7(3):e1001105.
8. Steinberg GB, Church BW, McCall CJ, Scott AB, Kalis BP. Novel predictive models for metabolic syndrome risk: a "big data" analytic approach. *Am J Manag Care*. 2014;20(6):e221-e228.
9. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. 2006. Available at: [http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes\\_new.pdf](http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf). Accessed June 23, 2015.
10. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2014;37(suppl 1):S81-S90.
11. Healthcare Cost and Utilization Project (sponsored by the Agency for Healthcare Research and Quality). Clinical Classifications Software (CCS) 2014. 2014. Available at: <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf>. Accessed June 23, 2015.
12. Toh S, Garcia Rodriguez LA, Hernan MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 2):13-20.
13. Friedman N, Koller D. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*. 2003;50:95-125.
14. Drew BG, Rye KA, Duffy SJ, Barter P, Kingwell BA. The emerging role of HDL in glucose metabolism. *Nat Rev Endocrinol*. 2012;8(4):237-245.
15. Lopez-Rios L, Novoa FJ, Chirino R, Varillas F, Boronat-Cortes M, Wagner AM. Interaction between cholesteryl ester transfer protein and hepatic lipase encoding genes and the risk of type 2 diabetes: results from the Telde study. *PLOS ONE*. 2011;6(11):e27208.
16. Barter PJ, Rye KA, Tardif JC, et al. Effect of torcetrapib on glucose, insulin, and hemoglobin A1c in subjects in the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial. *Circulation*. 2011;124(5):555-562.
17. Vozarova B, Stefan N, Lindsay RS, et al. High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes*. 2002;51(6):1889-1895.
18. Kunutsor SK, Apekey TA, Walley J. Liver aminotransferases and risk of incident type 2 diabetes: a systematic review and meta-analysis. *Am J Epidemiol*. 2013;178(2):159-171.
19. Marchesini G, Brizi M, Bianchi G, et al. Nonalcoholic fatty liver disease: a feature of the metabolic syndrome. *Diabetes*. 2001;50(8):1844-1850.
20. Dehghan A, van Hoek M, Sijbrands EJ, Stijnen T, Hofman A, Witterman JC. Risk of type 2 diabetes attributable to C-reactive protein and other risk factors. *Diabetes Care*. 2007;30(10):2695-2699.
21. Sanchez-Alavez M, Tabarean IV, et al. Insulin causes hyperthermia by direct inhibition of warm-sensitive neurons. *Diabetes*. 2010;59(1):43-50.
22. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall; 2004.
23. Jaynes ET. Information theory and statistical mechanics. *Physical Review*. 1957;106(4):620-630.
24. Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. New York, NY: John Wiley; 1992.
25. Schwartz GE. Estimating the dimension of a model. *Ann Statistics*. 1978;6(2):461-446.
26. Gull SF. Bayesian inductive inference and maximum entropy. In: Erickson GJ, Smith CR, eds. *Maximum Entropy and Bayesian Methods*. Dordrecht, Netherlands: Kluwer; 1988:53-74.
27. Heckerman D, Meek C, Cooper G. A Bayesian approach to causal discovery. Tech rep MSR-TR-97-05. Microsoft Research, 1997.
28. Madigan D, Raftery E. Model selection and accounting for model uncertainty in graphical models using occam's window. *J Amer Stat Assoc*. 1994;89:1535-1546.
29. Madigan D, York J. Bayesian graphical models for discrete data. *Int Stat Rev*. 1995;63:215-232.
30. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equation of state calculation by fast computing machines. *J Chem Phys*. 1953;21:1087-1092.
31. Robert C, Casella G. *Monte Carlo Statistical Methods*. New York, NY: Springer; 1989.