Title: **Development and validation of a multivariable Prediction Model for Pre-diabetes and Diabetes using Easily Obtainable Clinical Data**

Authors: Alan L. Hutchison, MD, PhD*[1], Mary E. Rinella, MD[1], Raghavendra G. Mirmira, MD, PhD[2], William F. Parker, MD, PhD[3]
*Alan.Hutchison@uchicagomedicine.org
1. Section of Gastroenterology, Hepatology and Nutrition, 2. Section of Endocrinology, Diabetes and Metabolism, 3. Section of Pulmonary and Critical Care, Department of Medicine, University of Chicago Medicine, Chicago, IL, 60637

Key Points

Question Can readily-available clinical data improve diagnosis of pre-diabetes and diabetes compared to hemoglobin A1c testing alone?

Findings In this cross-sectional study of 13,800 adults with paired hemoglobin A1c and oral glucose tolerance testing in the National Health and Nutrition Examination Survey, the rate of pre-diabetes undiagnosed by 8.6% and rate of diabetes undiagnosed by the hemoglobin A1c was 3.5%. A novel multivariable prediction model that included fasting plasma glucose, insulin, basic body measurements, and routinely available dyslipidemia and hepatic function labs for was significantly more accurate (AUROC 0.66/0.71 to 0.77 for pre-diabetes, 0.87/0.88 to 0.91 for diabetes) than hemoglobin A1C or fasting plasma glucose alone.

Meaning Incorporation of easily obtainable clinical data can improve diagnosis of pre-diabetes and diabetes compared to hemoglobin A1C alone.

Abstract

Importance In the US, pre-diabetes and diabetes are increasing in prevalence alongside other chronic diseases. Hemoglobin A1c is the most common diagnostic test for diabetes performed in the US, but it has known inaccuracies in the setting of other chronic diseases.

Objective To determine if easily obtained clinical data could be used to improve the diagnosis of pre-diabetes and diabetes compared to hemoglobin A1c alone.

Design, Setting, and Participants This cross-sectional study analyzed nationally representative data obtained from six 2-year cycles (2005 to 2006 through 2015 to 2016) of the National Health and Nutrition Examination Survey in the US. We excluded participants without hemoglobin A1c, oral glucose tolerance test, or sample weight data. The sample comprised 13,800 survey participants. Data analyses were performed from May 1, 2024 to February 9, 2025.

Main Outcomes and Measures We estimated 2-hour glucose from a gradient boosted machine decision tree machine learning model to diagnose pre-diabetes and diabetes as defined by oral glucose tolerance test 2-hour glucose of greater than or equal to 140 mg/dL but less than 200 mg/dL and greater than or equal to 200 mg/dL, respectively. We compared the area-under-the-receiver-operating-curve (AUROC), the calibration, positive predictive value, and the net benefit by decision curve analysis to hemoglobin A1C alone.

Results A 20-feature Model outperformed the hemoglobin A1c and fasting plasma glucose for diagnosis, with AUROC improvement from 0.66/0.71 to 0.77 for pre-diabetes and from 0.87/0.88 to 0.91 for diabetes. The Model also had improved positive predictive value compared to the A1c for diagnosis and for net benefit on decision curve analysis. Main features that improved diagnosis of pre-diabetes and diabetes were the standard vitals: age, height, weight, waist circumference, blood

pressure, pulse, the fasting labs plasma glucose, insulin, triglycerides, and iron, the non-fasting labs cholesterol, gamma-glutamyl transferase, creatinine, platelet count, segmented neutrophil percentage, urine albumin, and urine creatinine, and the social determinant of health factor Poverty Ratio.

Conclusions and Relevance In this cross-sectional study of NHANES participants, we identified risk factors that could be incorporated into the electronic medical record to identify patients with potentially undiagnosed pre-diabetes and diabetes. Implementation could improve diagnosis and lead to earlier intervention on disease before it becomes severe and complications develop.

Introduction:

The diabetes epidemic continues to grow in the United States, affecting an estimated 38.4 million people with 8.7 million likely undiagnosed based on epidemiological studies (1). Diabetes generates health care costs of $412.9 billion per year (2). As the diabetes epidemic has grown, other chronic diseases have also grown in prevalence, including obesity (35.7% in 2010 to 41.9% in 2022), kidney disease (13% in 2010 to 14% in 2023), metabolic-dysfunction associated steatotic liver disease (MASLD) (inconsistently diagnosed in 2010 to 24% in 2022), and sleep disorders (38.9% in 2013 to 44.3% in 2016) (1, 3-5).

The American Diabetes Association (ADA) recommends that options to diagnose pre-diabetes include an A1c of greater than or equal to 5.7%, a fasting plasma glucose between 100 and 125, or a 2-hour 75g OGTT plasma glucose of greater than or equal to 140 mg/dL, A diagnosis of diabetes can be made with any of the following: a hemoglobin A1c of 6.5%, a fasting plasma glucose greater than or equal to 126 mg/dL on two occasions in the absence of hyperglycemic symptoms, a random fasting plasma glucose greater than or equal to 200 mg/dL, or a 2-hour 75g OGTT plasma glucose greater than or equal to 200 mg/dL be used to diagnose diabetes. Despite these options, the hemoglobin A1c has become the predominant method clinically, due its estimation of the past 3 months of glycemia and due to its clinical ease of use as it does not require fasting (6, 7). Fasting plasma glucose has been shown to be discordant with hemoglobin A1c (8), but its use as primary diagnostic criteria is understudied due to electronic medical records not including fasting status (9).

However, it has been established that the hemoglobin A1c is inaccurate in the setting of chronic diseases such as chronic kidney disease, heart failure, cirrhosis, and after solid organ transplant (10, 11). One driver of this is the shortening of the mean lifespan of the erythrocyte, which disrupts the physiology on which the accuracy of the A1c is based. Unfortunately, these diseases are worsened by diabetes, with higher rates of morbidity and mortality for individuals with these diseases and diabetes (12). Current recommendations suggest use of the oral glucose tolerance test in some of these patient populations (13), but those are time-intensive tests that require fasting and a separate visit, and when applied at the population-level lead to high rates of non-adherence (14, 15). This gap puts these patients already suffering from chronic disease at risk of non-treatment for pre-diabetes and diabetes, for which we now have multiple treatment modalities including behavioral, technological, and pharmaceutical.

These gaps in the ability of the hemoglobin A1c are <u>well-understood by diabetologists</u>. However, from a public health standpoint, most <u>diabetes screening is not performed by diabetologists</u>, but instead by general practitioners, who, given the time demands on their practice (16), have to make efficient choices in their screening practices. Additionally, sending all their patients with the potential for inaccurate hemoglobin A1c labs is infeasible and unlikely to work given low adherence to oral glucose tolerance testing (14, 15).

One potential approach to improve diagnosis of pre-diabetes and diabetes is clinical incorporation of risk scores. Previously, incorporating additional risk factors into predictions of diabetes was laborious

and had to be performed separately patient by patient. The Cambridge Diabetes Risk Score and Bang Risk score incorporate factors such as age, gender, BMI, blood pressure medication, smoking history, and physical activity to predict risk of diabetes (17, 18). However, with the widespread adoption of the electronic medical record in the past decade, these features, and many more features, are available within these systems.

In the past decade computational advances have allowed for machine learning to aid in feature selection in large datasets. Two significant problems that plague large datasets are collinearity and missing data, which usually must be addressed with careful exclusionary variable selection and imputation of missing data. A powerful approach, decision tree analysis, bypasses both these concerns. This approach splits the data into separate branches at decision points (*e.g.* age greater than 60 years), combines variables into the same branch (*e.g.* left leg length and right leg length), and tests thousands of parameters to obtain the ideal set of parameters that predicts an outcome of interest the most efficiently and accurately. The premiere decision tree algorithm is a gradient-boosted decision tree analysis, XG Boost (19), which has been used to multiple outcomes, such as sepsis, heart failure mortality, and decompensated cirrhosis (20-22). This provides a powerful and validated approach to apply to the diagnosis of pre-diabetes and diabetes. A limitation of this algorithm is it cannot extrapolate – it can only make predictions within the training data it knows – therefore large well-sampled databases, such as the National Health and Nutrition Examination Survey, are ideal for generating models using XG Boost.

The National Health and Nutrition Examination Survey (NHANES) is a two-year repeating national survey consisting of anthropometrics, imaging, interviews, and laboratory values for several thousand non-repeating U.S. residents per cycle. Thirteen thousand eight hundred individuals underwent both a two-hour oral glucose tolerance test and hemoglobin A1c testing from 2005-2016. The OGTT was performed on participants aged 12 and greater who were randomly assigned to present for a morning fasting examination. Exclusion criteria included being on anti-diabetes medications and refusal to adhere to the testing requirements. NHANES provides a high-fidelity and nationally representative data source to understand what readily available clinical measurements, laboratory values, patient history questions can be used to better estimate the two-hour glucose and identify undiagnosed diabetes. In addition to the OGTT and hemoglobin A1c, participants from 2005-2016 underwent additional separate testing for a series of fasting blood test, anthropometrics, and questionnaires. A strength of the NHANES database is that while there are niche tests obtained that are not regularly gathered during standard clinical care, most of the variables are ones that are easily obtained during routine adult internal medicine and therefore potentially accessible in the electronic medical record.

In this work, we aim to identify undiagnosed diabetes from easily obtained clinical data including the hemoglobin A1c, with the 2-hour OGTT as our gold-standard comparator. To do so, we apply a machine learning approach to the NHANES dataset from 2005-2016 and identify features that improve the predictive ability of our model in diagnosing pre-diabetes and diabetes.

Methods: We obtained data from the NHANES website, taking data from 18 surveys from 2005-2016 to capture a comprehensive set of risk factors including vital signs, anthropometrics, baseline labs, fasting labs, and additional labs (Supplemental Methods, Supplementary Table 2). Incorporating that data, we first excluded participants for whom we did not have hemoglobin A1c, two-hour glucose, and sample weighting data (Supplementary Figure 5). After this, to increase the predictive ability of our model, we excluded features that were missing in more than 12% of the data. Further variable stratification was done based on eliminating overlapping variables (Supplementary Table 3). We then divided our data into the training dataset and evaluation dataset in accordance with the TRIPOD-AI guidelines (23), using data from 2005-2014 for the training data and data from 2015-2016 for the evaluation data to predict the two-hour glucose. We used gradient boosted decision trees in *R* using the

xgboost package and *caret* for hyperparameter tuning. The *AUC* and *pROC* packages were used for confusion matrix calculations, the *glm* package was used for binomial modeling, and the *dcurves* package was used for Decision Curve Analysis modeling. Plotting was done using *cowplot* and *ggplot2. R* Markdown files and data are available at the *GitHub* repository *alanlhutchison/NHANES-DMDx*, and the raw data is available from the Center for Disease Control and Prevention.

We generated a Full model using all the available variables as above but then picked the 20 most important features to generate a minimal model, which we refer to as the Model.

Table 1: Participant Demographic Features

	2005-2014	**2015-2016**
Fasting Plasma Glucose	101.8 (±18.0)	102.6 (±18.9)
Hemoglobin A1c	5.4 (±0.6)	5.5 (±0.6)
Age	41.1 (±20.6)	41.8 (±20.3)
Insulin		
Mean (SD)	10.9 (±9.6)	12.4 (±12.1)
Missing	232 (2.0%)	32 (1.5%)
Triglycerides		
Mean (SD)	115.0 (±100.4)	115.5 (±95.2)
Missing	91 (0.8%)	6 (0.3%)
Height		
Mean (SD)	167.4 (±10.1)	166.3 (±10.1)
Missing	73 (0.6%)	7 (0.3%)
Waist Circumference		
Mean (SD)	94.4 (±16.7)	95.9 (±17.5)
Missing	189 (1.6%)	41 (2.0%)
Systolic Blood Pressure		
Mean (SD)	119.2 (±16.8)	121.7 (±17.5)
Missing	258 (2.2%)	26 (1.2%)
GGT		
Mean (SD)	25.9 (±37.3)	25.1 (±31.1)
Missing	83 (0.7%)	6 (0.3%)
Cholesterol		
Mean (SD)	187.9 (±41.6)	188.9 (±42.7)
Missing	83 (0.7%)	6 (0.3%)
Urine Albumin		
Mean (SD)	28.3 (±174.5)	29.8 (±159.6)
Missing	72 (0.6%)	11 (0.5%)
% Segmented Neutrophils		
Mean (SD)	56.1 (±9.9)	55.8 (±9.4)
Missing	43 (0.4%)	9 (0.4%)
Creatinine		
Mean (SD)	0.8 (±0.3)	0.8 (±0.3)
Missing	81 (0.7%)	6 (0.3%)
Urine Creatinine		
Mean (SD)	129.0 (±78.0)	129.0 (±81.2)
Missing	72 (0.6%)	11 (0.5%)
Pulse		
Mean (SD)	72.1 (±11.8)	72.3 (±11.4)
Missing	220 (1.9%)	25 (1.2%)
Poverty Ratio		
Mean (SD)	2.5 (±1.6)	2.4 (±1.6)
Missing	817 (7.0%)	193 (9.3%)
Diastolic Blood Pressure		
Mean (SD)	67.1 (±11.9)	67.2 (±11.6)
Missing	309 (2.6%)	32 (1.5%)
Platelets		
Mean (SD)	251.4 (±66.4)	237.5 (±58.8)
Missing	22 (0.2%)	9 (0.4%)
Iron		
Mean (SD)	88.0 (±36.3)	85.5 (±35.5)
Missing	90 (0.8%)	6 (0.3%)
Weight		
Mean (SD)	77.8 (±21.5)	78.5 (±22.0)
Missing	67 (0.6%)	8 (0.4%)

Results:

Baseline characteristics:

Participants with missing hemoglobin A1c, 2-hour glucose, or sample weights were excluded from the analysis. Starting with 60,936 participants, excluding participants reduced the number to 13,800 participants (STROBE Diagram, Supplementary Figure 5). These were further separated by year of evaluation into 2005-2014 (training dataset, 11,718 participants) and 2015-2016 (evaluation dataset, 2,082 participants) cohorts. Table 1 shows the summary features for the cohorts, based on the top twenty features in the model described below.

To obtain a sample reflective of the US population, we resampled the 13,800 participants 100,000 times with replacement using the NHANES-assigned sample weights for their oral glucose tolerance test data.

Of the resampled population (which excluded those on anti-diabetes medication as per OGTT exclusion criteria (Supplementary Table 4)), 5.3% had a 2-hour glucose or hemoglobin A1c diagnostic of diabetes. 0.4% had a diagnostic hemoglobin A1c but a non-diagnostic 2-hour glucose (1.7% total had a diagnostic hemoglobin A1c), and 3.5% had a diagnostic 2-hour glucose but a non-diagnostic hemoglobin A1c (4.9% total had a diagnostic 2-hour glucose). For pre-diabetes, 29% had a 2-hour glucose or hemoglobin A1c diagnostic of pre-diabetes. 12.3% had a diagnostic hemoglobin A1c but a non-diagnostic 2-hour glucose (20.1% total had a diagnostic hemoglobin A1c), and 8.7% had a diagnostic 2-hour glucose but a non-diagnostic hemoglobin A1c (14.5% total had a diagnostic 2-hour glucose). Supplementary Table 5 and Supplementary Table 6 show the confusion matrix for the diagnosis of diabetes and pre-diabetes.

Of note – in this analysis we will focus on individuals who would be underdiagnosed as not having diabetes if assessed with the hemoglobin A1c alone, which is why we are using the 2-hour glucose as the gold standard. The diagnosis of diabetes can be done with either, but those who have been diagnosed with the hemoglobin A1c are not those a risk of a missed diagnosis.

Association of A1c with 2-hour glucose in patients with A1c < 6.5%

There are several diseases in which we know the A1c is less accurate, and several conditions that increase the risk of diabetes – given then, we wanted to assess how the relationship between 2-hour glucose and A1c changed under these chronic conditions for hemoglobin A1c < 6.5%. Given high rates of missing data in the medical history questionnaires in NHANES, we focused on lab values reflective of diseases (Figure 1). For chronic kidney disease, we chose creatinine. For liver disease, we chose the non-invasive fibrosis test FIB-4. For anemia, which can impact hemoglobin A1c levels, we chose hemoglobin. For obesity, we used the BMI threshold of 30. For liver disease, we used the non-invasive test for fibrosis, the FIB-4 (24). For age, we used a threshold of 60 years. For hypertension we focused on the systolic blood pressure. Though the resampling precludes use of statistical significance testing, it is clear from visual inspection that higher creatinine, BMI, FIB-4, age, and systolic blood pressure all lead to higher relative two-hour glucose for a given hemoglobin A1c range.

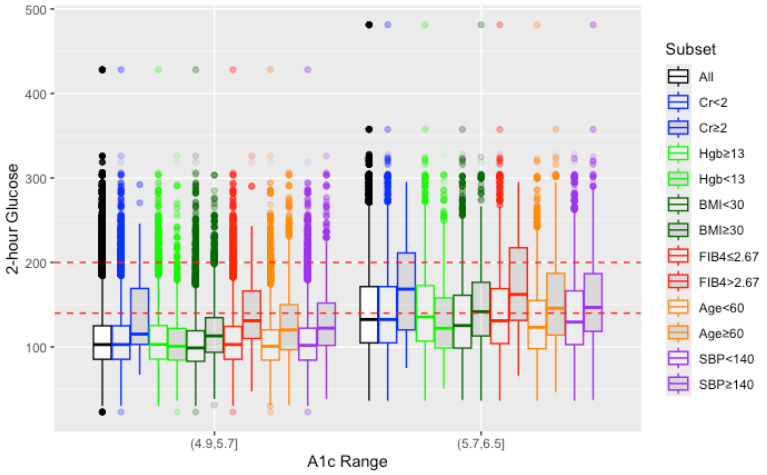


Figure 1. Relationship between Hemoglobin A1c and two-hour glucose stratified by A1c range. Red lines show thresholds used to diagnose pre-diabetes (140 mg/dL ≤2-hour glucose < 200) and diabetes (2-hour glucose ≥ 200). Data from NHANES 2005-2016 with hemoglobin A1c and two-hour glucose resampled with replacement 100,000 times based on OGTT sample weights.

Top features of interest in the Model

The resulting model incorporated 78 variables (Full Model). The top 20 variables were taken to define the Minimal Model.

The importance of the top 20 features in the Full Model is shown in Supplementary Figure 6. The laboratory values that are affected by fasting (fasting plasma glucose, insulin, and triglycerides) have the largest predictive effect, ranking 1, 4, and 5 respectively in the Full model. Clinically-accessible vital signs included age, height, waist circumference, systolic blood pressure, pulse, diastolic blood pressure, and weight. Non-fasting labs that were the most important included the hemoglobin A1c, gamma-glutamyl transferase, total cholesterol, urine albumin, percentage segmented neutrophils, urine creatinine, creatinine, platelets, and iron. The Poverty Ratio of the zip code of the participant was also in the top 20 variables.

We compared the results of our Model-predicted two-hour glucose to the meausred two-hour glucose, shown in Figure 2. The R^2 for the fit between the predicted and measured values was 0.62.

Comparison with the hemoglobin A1c for diagnosing pre-diabetes and diabetes

We compared the Full model to the minimal model (referred to as the Model) and found that there was minimal difference in test characteristics (Supplementary Figure 7, Supplementary Table 7) allowing us

to focus on comparison between the Model and the A1c (Figure 3). We found that the Model had superior AUROC compared to the A1c and FPG for pre-diabetes (Model: 0.77 vs. A1c: 0.66 vs FPG: 0.71) and for diabetes (Model: 0.91 vs. A1c: 0.87 vs FPG: 0.88).

Additionally, we trained our Full and Minimal Models (Supplementary Figure 8, Supplementary Table 7) on cohorts with co-morbid chronic diseases – Age > 60, BMI > 30, eGFR < 60, FIB-4>2.67, and Systolic BP > 140. We found that for these cohorts, the Model outperformed the A1c in all cases for diagnosing pre-diabetes, though in some cases both the Model and the A1c performed poorly. For diagnosing diabetes, the model performed better in the cases of BMI > 30, equivalent by AUROC for FIB-4 > 2.67, Age > 60, SBP > 140, and eGFR < 60. The FPG performed more similarly to the hemoglobin A1c than to the model, though notably it performed better than the hemoglobin A1c and Model for identifying pre-diabetes in participants with eGFR < 60 and diabetes in participants with FIB-4 > 2.67. A major limit of this exploratory analysis, however, is the low sample size in the evaluation cohort for these chronic disease cohorts. Given the comparable behavior of the hemoglobin A1c and the FPG, in the subsequent analysis we focus on comparison of the Model with the hemoglobin A1c.

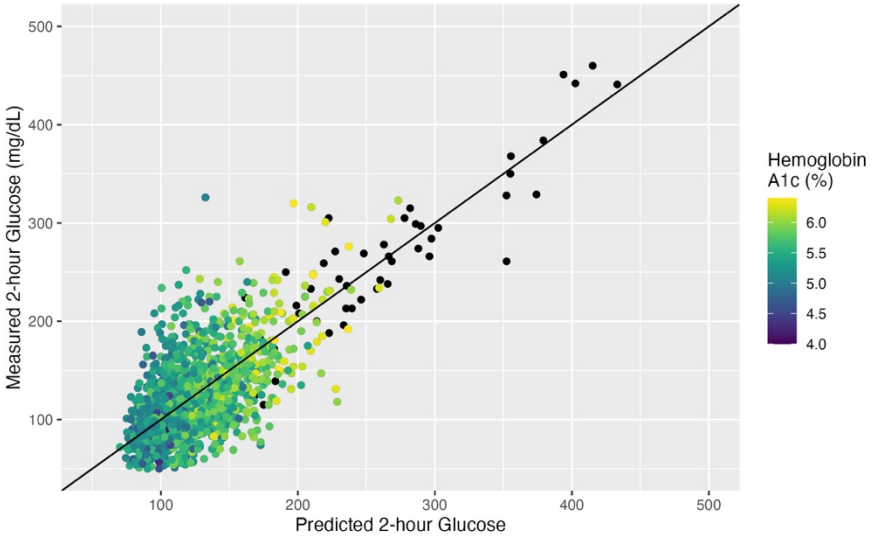


Figure 2. Calibration plot of Model-predicted 2-hour glucose (x-axis) compared to measured 2-hour glucose (y-axis). Points are colored for hemoglobin A1c < 6.5% (legend), black for hemoglobin A1c > 6.5% to aid with visualization.

We wanted a parallel way of establishing the utility of our models for predicting undiagnosed pre-diabetes and diabetes. Decision Curve Analysis measures the net benefit difference between a set of tests or interventions as the threshold at which a patient or provider would want to receive the test based on their risk of the diagnosis changes (Figure 4). We found that for a diagnosis of pre-diabetes, among those who would want an additional test if they had a 15-20% risk of pre-diabetes, there was net benefit of 0.0375 to using the Model over the A1c as a test of pre-diabetes versus performing no testing at all – that is, if 1000 people who would want the test if they had a 15-20% risk of pre-diabetes received it, 37 would benefit from the Model over using the A1c alone. For a diagnosis of diabetes, the net benefit above 25% is 0.005, that is, 5 out of 1000 would benefit from the Model over the A1c alone. Supplementary Figure 9 and Supplementary Figure 10 show the decision-curve analysis for the different cohorts for pre-diabetes and diabetes.

<u>Positive Predictive Values for Diagnosis</u>

We wanted to understand if the A1c could help us determine when we would benefit from using the Model to determine when we might need an oral glucose tolerance test to diagnose pre-diabetes or diabetes.

To accomplish this, we additionally examined our resampled data to assess the predictive value of pre-diabetes and diabetes for different A1c values. In Supplemental Figure 11 we stratified resampled participants by their two-hour glucose-based diagnosis and looked to see how the Model and A1c

classified their positive predictive value (PPV) of having diabetes for participants with A1c < 6.5%. The blue lines indicate the PPV by the A1c – it includes no other information so is uniform across the stratification groups. The PPV of the model is represented as boxplots for each group and value of A1c (A1c < 5.7 are binned together). For all values of A1c < 6.4%, the Model had a superior median PPV to the A1c for predicting diabetes in participants with diabetes diagnosed by 2-hour glucose. Additionally, the Model had lower PPV than the A1c for diabetes in participants with A1c 5.9 to 6.4 and pre-diabetes on 2-hour glucose.

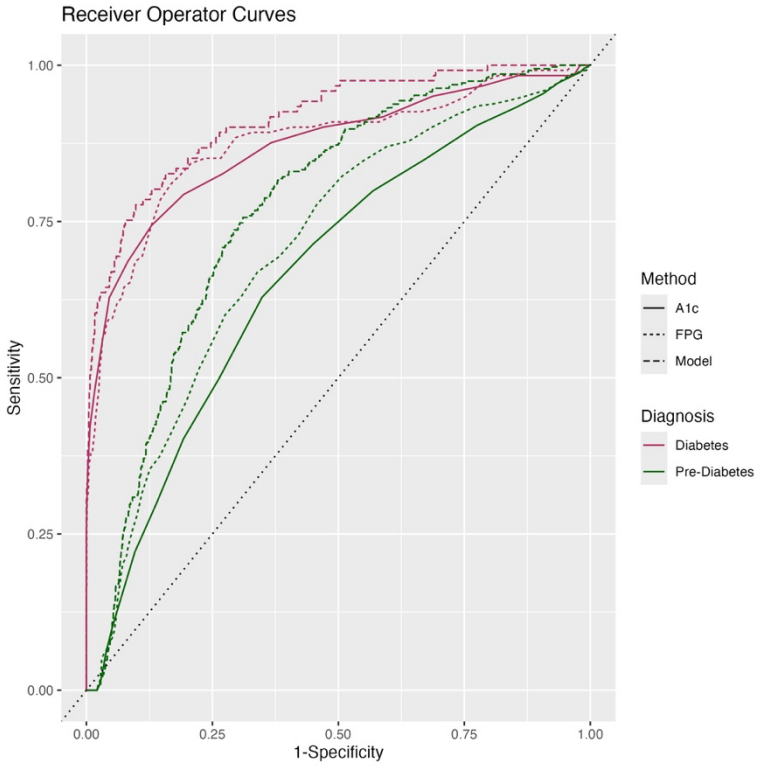


Figure 3. ROC comparing A1c to Model of 20 most important features.

To understand our data in a different way, we examined how using the Model instead of the A1c changed the probably of diabetes for the groups stratified by 2-hour glucose for participants with A1c < 6.5%. Supplemental Figure 12 shows that both methods miss some participants (see top purple band across top of Supplemental Figure 12), more participants with 2-hour glucose-diagnosed diabetes are correctly re-classified into higher PPV categories (more purple goes toward the bottom of the right column, while more blue goes towards the top of the right column). The stacks are defined as greater than 70% PPV (high likelihood), 39-70% PPV (intermediate likelihood), 6-39% PPV (more likely that participants have pre-diabetes over diabetes (Supplementary Figure 13) and less than 6% PPV, which is the basal rate of diabetes for this resampling.

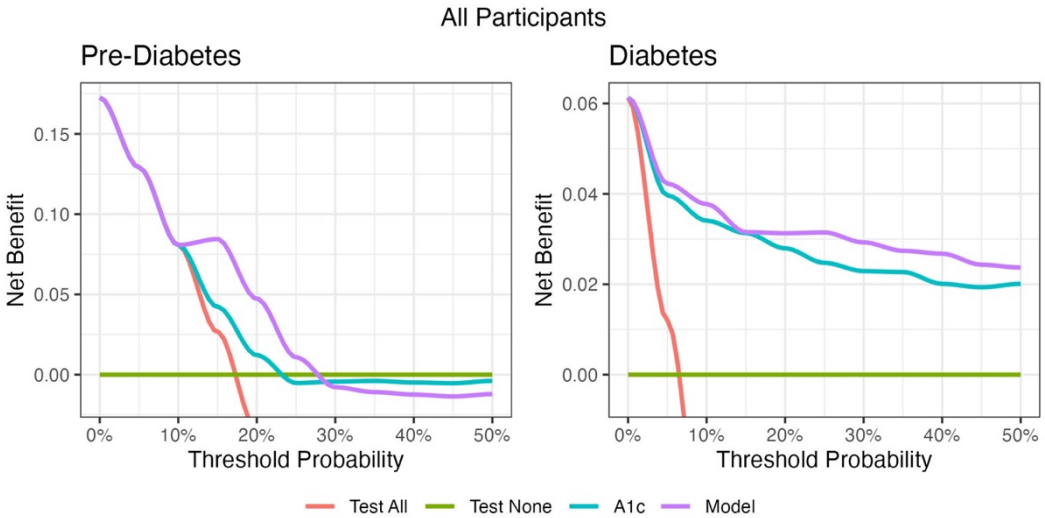


Figure 4. Using Decision Curve Analysis, the Evaluation data shows that the Model outperform the A1c for predicting two-hour glucose-defined diagnosis of pre-diabetes (left panel) and diabetes (right panel). The x-axis indicates the probability of a diagnosis needed to consider testing, and the y-axis shows increased benefit using a given strategy for testing – here, greater benefit using the Model to predict who has diabetes who may be missed by the A1c.

While maximizing AUC and PPV is important in a model, we also wanted to understand the contribution of true positives, false positives, true negatives, and false negatives to this accuracy. We resampled our data 10,000 times with replacement with the OGTT sample weights, then stratified our

data by A1c range for pre-diabetes and diabetes (Supplementary Figure 14 and Supplementary Figure 15). We found that the Model results in fewer false positives than the A1c for pre-diabetes, but more false positives for diagnosis of diabetes, though also with fewer false negatives.

Discussion

In this work, we developed a machine learning-based model that allows for personalization of identification of pre-diabetes and diabetes in patients who would not otherwise be identified as at risk. We established that even with commonly obtained laboratory values (Complete Blood Count with Differential, Comprehensive Metabolic Panel, hemoglobin A1c, gamma-glutamyl transferase, total cholesterol, urine studies), a few additional fasting studies (plasma glucose, insulin, triglycerides, iron), clinic vital signs, and a questions regarding past medical history, family history, and social determinants of health, the ability to predict pre-diabetes and diabetes can be greatly improved over using the hemoglobin A1c alone.

Of the features identified by our models, many make clear physiological sense. Of the vital sign features, age, height, weight, waist circumference, heart rate, and blood pressure all have strong associations with diabetes, insulin resistance, and metabolic syndrome. Other than the hemoglobin A1c, the most important labs were the fasting plasma glucose, insulin, triglycerides, and total cholesterol, which is already linked to diabetes. The creatinine and urine albumin-creatinine ratio are important markers of diabetic nephropathy, supporting their identification by the model.

There were findings from our model that are associated with insulin resistance and diabetes but less expected to be at the top of our model. The percentage of segmented neutrophils was identified as a key feature – high absolute neutrophil counts have been associated with diabetes and worse cardiovascular outcomes (25). Platelet dysfunction, aggregation, and mean volume have been associated with diabetes, but platelet levels themselves have been more weakly correlated (26), though they can be a marker of liver disease. Iron levels are associated with overall nutritional status, chronic disease, and chronic inflammation, all of which can play a role in insulin resistance and diabetes. Gamma-glutamyl transferase is specific for liver and has previously been associated with diabetes both in the NHANES dataset (27) and independently (28). Of the SDOH history, Poverty Ratio was the only feature in the top 20 (#16). Poverty is well-established to increase type 2 diabetes incidence even in the setting of universal health care (29, 30), though its use diagnostically has not been noted.

There are other risk scores for predicting and identifying diabetes: the FINDRISC score uses age, gender, BMI, use of blood pressure medication, a family history diabetes, history of high blood glucose, physical activity, and daily consumption of vegetables, fruits, or berries (31). The Cambridge Diabetes Risk Score includes age, gender, BMI, use of blood pressure medication, prescribed steroids, and smoking history (17). A patient questionnaire-based approach using earlier NHANES data (1999-2006) as well as 2 community studies found similar results of age, sex, family history of diabetes, history of hypertension, obesity, and physical activity (18). Our model captured many of these features but excluded some that had too much missing data in the NHANES database. Importantly, our model focuses on individuals for whom the A1c is non-diagnostic of diabetes and only included those not already on diabetes medications (as those participants did not receive oral glucose tolerance testing), providing a tool for assessing a population previously thought to not be at risk.

In addition to having diagnostic importance, the hemoglobin A1c is also used as a therapeutic goal, both for individual patient management and for health care system quality metrics. However, a decreasing hemoglobin A1c is often interpreted as “improving” even though there are several aspects of chronic disease that can decrease the A1c. In cirrhosis and chronic kidney disease, red blood cell turnover time can decrease the A1c artificially, as the hemoglobin molecules have less time to be exposed to hyperglycemia. Identification of patients with diabetes for whom the hemoglobin A1c is

already below “goal levels” may pose a therapeutic challenge for providers. Alternative methods, such as self-monitoring of glucose via capillary blood glucose monitoring, continuous glucose monitoring, or serial oral glucose tolerance tests may be warranted to assess therapeutic benefits of interventions.

In our analytical approach we attempted to avoid turning continuous variables into ordinal or categorical ones as much as possible (32). Given that our goal is matching the clinical diagnosis of diabetes, which requires a definition using a clear cutoff of 2-hour glucose of 200 mg/dL or A1c ≥ 6.5%, we had to dichotomize our analysis for this purpose. As we have demonstrated, risk for diabetes does not match a sharp biomarker cutoff.

There is a well-established role that sleep quality plays in diabetes risk and progression. However, in our model the self-reported sleep time was a minimal contributor to the risk of diabetes. It is likely that to be a more significant predictor of diabetes a more intensive sleep assessment would be needed.

What new clinical paradigms does our model suggest for the clinician? Though diabetologists treating people diagnosed with diabetes are well-versed in the limitations of the hemoglobin A1c and have resources available to arrange, obtain, and encourage adherence to other modalities of management, the general practitioner unlikely has these resources or time to dedicate. Given the growing burden of chronic diseases other than diabetes, new paradigms are needed to screen for diabetes. The results of this modeling suggest that with the simple addition of fasting labs and urine testing we could greatly increase our detection of undiagnosed pre-diabetes and diabetes. Patients being screened for pre-diabetes or diabetes should likely be receiving screening for metabolic-dysfunction associated steatotic liver disease (MASLD) as well, which requires a CBC and CMP at baseline already (24). Cholesterol screening is already recommended for adults who are 40 years old or older by the US Preventive Service Task Force and American College of Cardiology/ American Heart Association (33, 34). The additional medical history, family history, and social determinants of health questions can be easily filled out based on a questionnaire in clinic, through the electronic medical record, by medical support staff and should not add to clinical burden. Patients whom models like this identify as at risk of pre-diabetes or diabetes could then be referred for a second round of testing, involving fasting plasma glucose, fasting insulin, fasting triglycerides, fasting iron, and urine albumin and urine creatinine. After this, patients for whom the model suggests a diagnosis of pre-diabetes or diabetes may benefit first from prompt re-evaluation and then with referral for an oral glucose tolerance test. Though this is more complicated than a simple non-fasting blood test such as the hemoglobin A1c in isolation, with the multiple chronic diseases that need to be screened for and the integration of the electronic health record, this approach should benefit patients without additionally burdening clinical care.

Limitations

This study has limitations. The NHANES data used is from 2005-2016 – since then, rates of overweight, obesity, chronic kidney disease, and MASLD have increased even further, likely increasing the rates of diabetes even further. These measurements of hemoglobin A1c and OGTT represent only a snapshot in time – oral glucose tolerance test results have been shown to be influenced by time of day as well as by the prior day’s diet and sleep (35-37), which could cause both false positive and false negative results – though these effects may average out on a population level. Information that may have provided more information, such as body composition assessment by DEXA scan, which has implications for peripheral insulin resistance, had too much missing data to be used for our purposes.

The participants in NHANES who obtained oral glucose tolerance testing were a subsample of the whole set of participants. Exclusion criteria are detailed in the STROBE diagram (Supplementary Figure 5) and Supplementary Table 4. The main exclusion criteria other than older than 11 years of age was not being on anti-diabetes medications (even if there was a diagnosis of diabetes). Participants were further randomly assigned to a morning exam and fasting. A small number refused to fast, or did not

fast long enough, or refused to adhere to some other part of the exam. Despite this, we believe that this subsample (with separate sampling weights designated by NHANES) is representative of the larger US population.

The fasting plasma glucose is a separate modality for diagnosing diabetes – in the absence of hyperglycemic symptoms this requires two separate readings – the NHANES database we used is not longitudinal, limiting the use of the fasting plasma glucose as a diagnostic marker, though we show that our model outperforms a single fasting level. Our model does identify it as one of the important fasting labs – along with insulin, triglycerides, and iron. Insulin is not obtained in standard clinical care, and historically it was very difficult to measure accurately. The use of the HOMA-IR as a screening tool has been suggested (38), and requires further consideration for cost-analysis, as insulin is not otherwise regularly collected.

Developing models and diagnostic tests always requires a balance between false negative results and false positive results. We have demonstrated that using the A1c exclusively leads to many false negatives. However, all the models we generated came with a degree of false positive results. (Supplementary Figure 14 and Supplementary Figure 15). These are cases that the model may have indicated someone needed further testing which would have ultimately been normal.

Further studies are needed to validate this work. All this work is reproducible as described above. Validation of this model in other datasets with paired hemoglobin A1c and oral glucose tolerance testing would be warranted, as would validation of this prospectively, first in a single health system and eventually in a multi-site prospective study. Cost-effectiveness evaluation would also be an important feature of future work.

Conclusion

In this work, we have established the ability of easily accessible labs and questionnaires to better risk-stratify patients with undiagnosed diabetes who would benefit from oral glucose tolerance testing. Our machine learning approach allows for personalization of identification of pre-diabetes and diabetes in patients who would not otherwise be identified as at risk. Given the single-database single-timepoint-per-patient nature of this analysis, validation in external datasets will be needed. Future work is needed to validate and refine these findings, improving our ability to personalize diabetes risk stratification for patients.

Supplemental Methods

We included the surveys *Demographics (DEMO), Alcohol Use (ALQ), Blood Pressure (BPQ)*, *Diabetes (DIQ)*, *Sleep Disorders (SLQ),* and *Medical Conditions (MCQ) Questionnaires.* We additionally included *Body Measurements (BMX), Blood Pressure (BPX)*, *Complete Blood Count (CBC), Biochemistry Profile (BIOPRO), Glycohemoglobin (GHB)*, *Magnesium and Related Tests (MGX), Dual-Energy X-ray Absorptiometry (DXXAG),* Fasting Questionnaire (*FASTQX), Hepatitis A (HEPA, Hepatitis B (HEPB_S), Hepatitis C (HEPC),* and *Hepatitis E (HEPE)* data. We adjusted lab values to 2015-2016 measurements in accordance with NHANES recommendations based on changes in methodology and laboratory measurement techniques.

In total, the variables include the following: basic information: vital signs collected in clinic, demographics, complete blood count, comprehensive metabolic panel, and hemoglobin A1c; fasting labs: fasting plasma glucose, fasting triglycerides, fasting insulin, and fasting iron; social determinants of health questionnaire; additional labs that may not be as frequently obtained in routine clinical care: the urine labs urine albumin and urine creatinine, the blood labs phosphorus, total cholesterol, gamma-glutamyl transferase (GGT), Hepatitis B core and surface antibody, globulin, lactate dehydrogenase, hydroxycontine, hepatitis A total antibody, and osmolality.

Acknowledgements
We would like to thank Dr. Esra Tasali, Dr. Celeste Thomas, and Dr. Brandon Faubert for their feedback in preparation of this manuscript. No patients were involved in this study design or preparation of this manuscript.

Funding
Dr. Hutchison is supported by NIH P30 DK020595, NIH T32 DK007074, and the American Association for the Study of the Liver. Dr. Mirmira is supported by NIH U01 DK127786, R01 DK060581, and R01 DK105588. Dr. Parker is supported by NIH K08HL150291 and R01LM014263 and the Greenwall Foundation. Dr. Hutchison, Dr. Parker, Dr. Mirmira, and Dr. Rinella have no relevant financial disclosures.

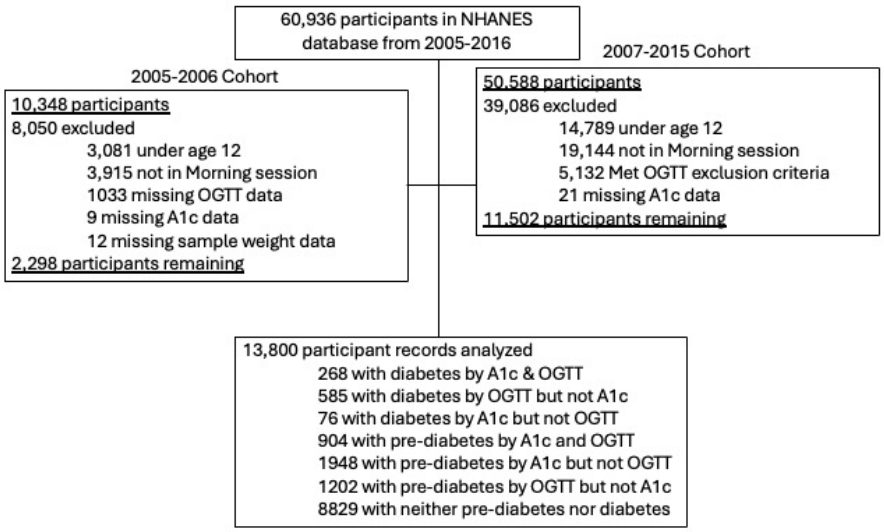
Supplementary Figures and Tables

Supplementary Table 2. Features included in the machine learning Full model.

Feature Type	Feature
Obtained in Clinic	**Age**, Gender, **Pulse**, **Systolic BP, Diastolic BP**, **Height**, Weight, **Waist Circumference**, Arm Circumference, (Ir)Regularity of Heartbeat
History-obtained	Have you ever been told you had (High BP, Overweight, Pre-diabetes, Poor Sleep, Diabetes, Asthma) Have you ever been treated for Anemia Have you ever received a transfusion? How many hours of sleep do you have a night? Do you have a relative with asthma? Have you been tested for DM in the past 3 years?
CBC & Differential	Hemoglobin, White Blood Cell count, **Platelet count**, Mean Cell Volume, Mean Platelet Volume, Red Blood Cell Count, Mean corpuscular hemoglobin concentration, Percentage of (**Neutrophils**, Lymphocytes, Monocytes, Eosinophils, Basophils), Mean Cell Hemoglobin, RBC Distribution Width
Comprehensive Metabolic Panel + Hemoglobin A1c	Sodium, Potassium, Chloride, Bicarbonate, Blood Urea Nitrogen, **Creatinine**, Calcium, Total Protein, Total Bilirubin, Albumin, Aspartate aminotransferase, Alanine aminotransferase, Alkaline Phosphatase, **Hemoglobin A1c**
Fasting Labs	**Fasting Plasma Glucose, Fasting Triglycerides, Fasting Insulin, Fasting Iron**
Social Determinants of Health	**Poverty Ratio**, Race/Ethnicity, Number of People in Household, Marriage Status, Primary Language, Citizenship Status, Education Level
Urine Labs	**Urine albumin, Urine creatinine**
Various Labs	**Total cholesterol**, Phosphate, **Gamma-Glutamyl Transferase**, Hepatitis B surface Antibody, Hepatitis B core Antibody
	Uric Acid, Hydrocontine, Hepatitis A Antibody, Osmolality, Globulin, Lactic Dehydrogenase,

Supplementary Table 3: Redundant features excluded from analysis

Feature	Included Feature	Excluded Features
Blood pressure	Systolic Blood Pressure average	BP site, BP arm, BP cuff peak pressure, BP enhancement usage
Length	Height	Arm length, Upper leg length
Weight	Weight	BMI
Language	Language	Interpreter usage, Proxy usage
Family Size	Number in family	Number in household



Supplementary Figure 5 STROBE diagram.

Supplementary Table 4 OGTT participation or exclusion criteria for NHANES data collection years 2007-2015. Years 2005-2006 had similar exclusion criteria, but did not report the reasons for the exclusions as in subsequent years.

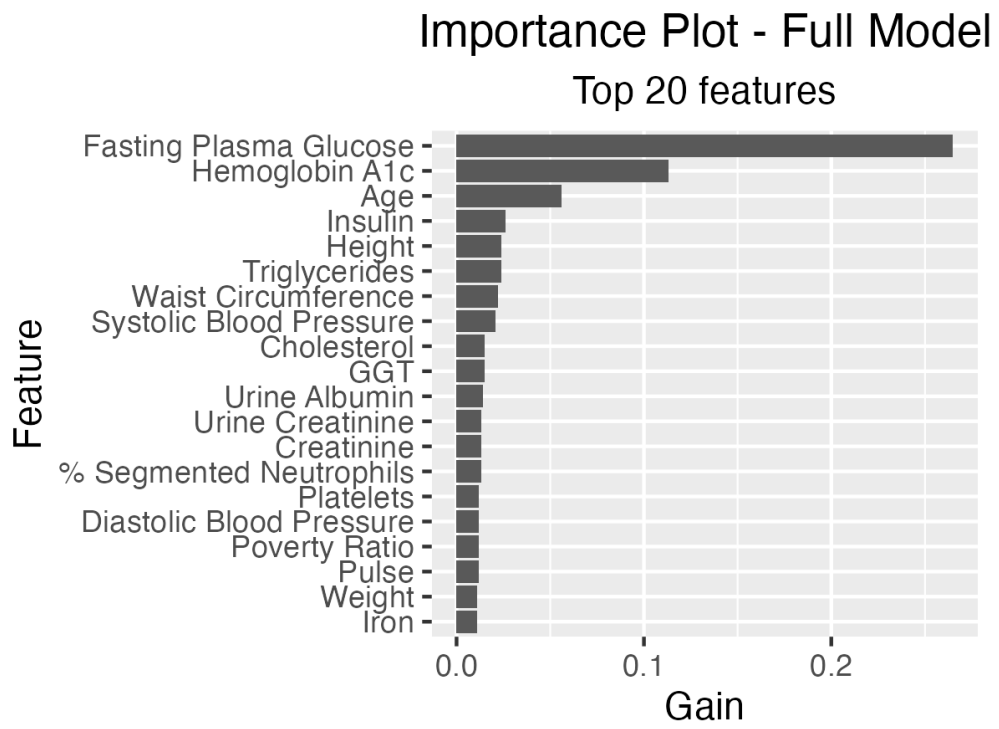
OGTT Examination	**Number**
OGTT Complete	11523
Hemophiliac	4
Cancer chemotherapy within 4 weeks	11
Diabetic on medications	1537
Refused venipuncture	58
Ill/faint during test	43
Venipuncture unsuccessful	84
Currently pregnant	25
Refused glucose challenge	293
Fasting less than 9 hours	323
Came late/left early	122
Other	892
OGTT omission information absent	1740
Total OGTT not completed	5132

Supplementary Table 5. Confusion matrix for the ability of the hemoglobin A1c diabetes thresholds to predict two-hour glucose diagnostic thresholds for diabetes.

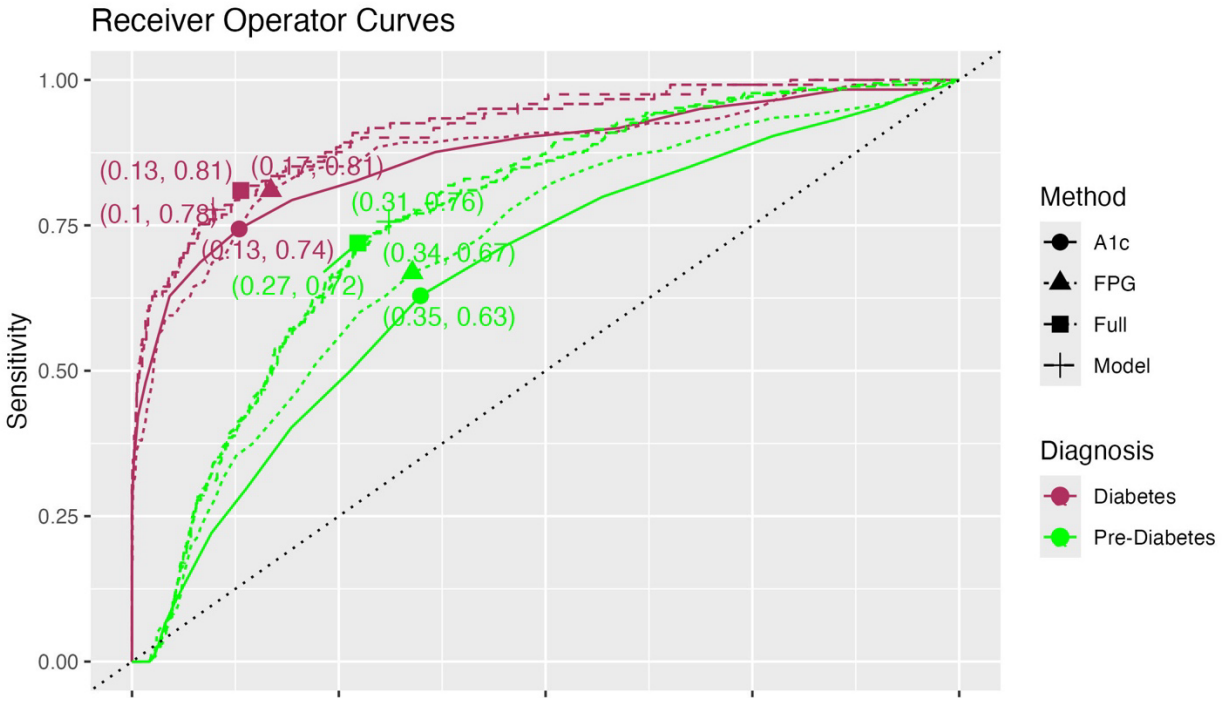
	A1c ≥ 6.5%	**A1c < 6.5%**	
2-hour Glucose ≥ 200	1.3%	3.5%	Sensitivity: 27.1%
2-hour Glucose < 200	0.4%	94.7%	Specificity: 99.6%
	PPV: 76.5%	NPV: 96.4%	

Supplementary Table 6. Confusion matrix for the ability of the hemoglobin A1c pre-diabetes thresholds to predict two-hour glucose diagnostic thresholds for pre-diabetes. PPV: Positive Predictive Value. NPV: Negative Predictive Value.

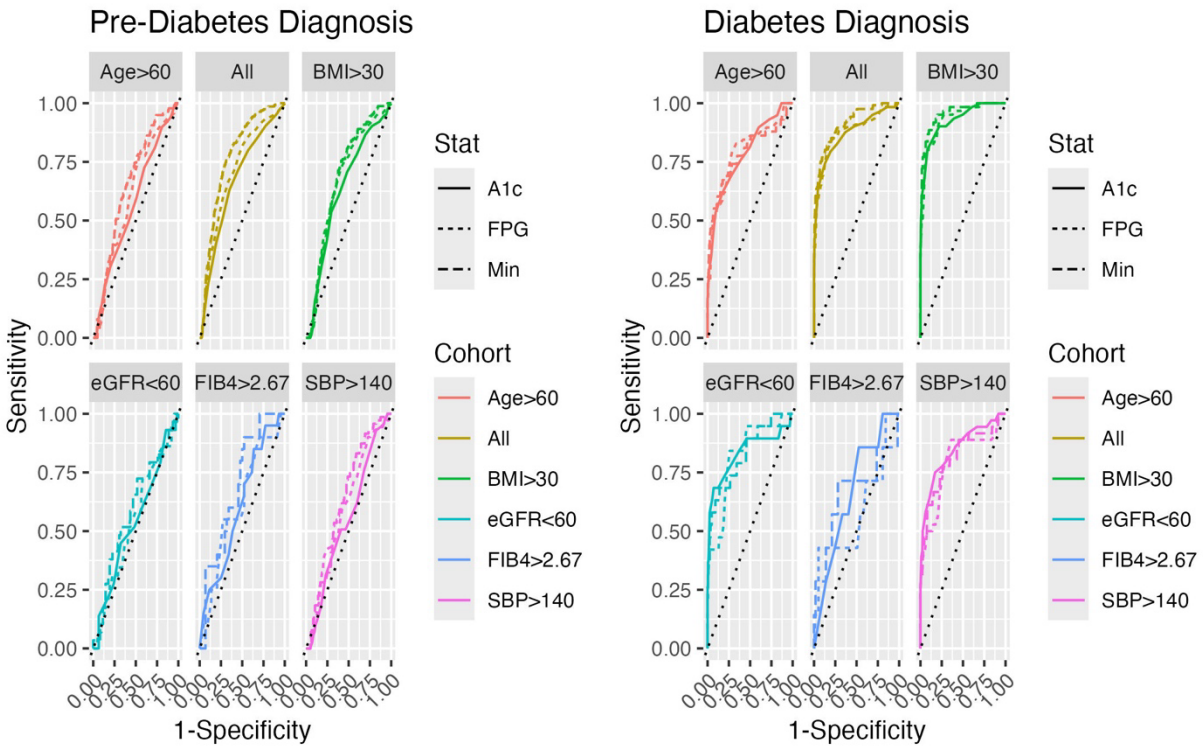
	5.7% ≤ A1c < 6.5%	**A1c < 5.7%**	
140 ≤ 2-hour Glucose < 200	29.0%	8.7%	Sensitivity: 76.9%
2-hour Glucose < 140	12.3%	50%	Specificity: 80.3%
	PPV: 70.2%	NPV: 85.2%	



Supplementary Figure 6. This importance plot highlights the variables in NHANES that are predictive of the two-hour glucose for the 5 different models. Only top 20 variables shown, though all eligible variables were included based on the model.



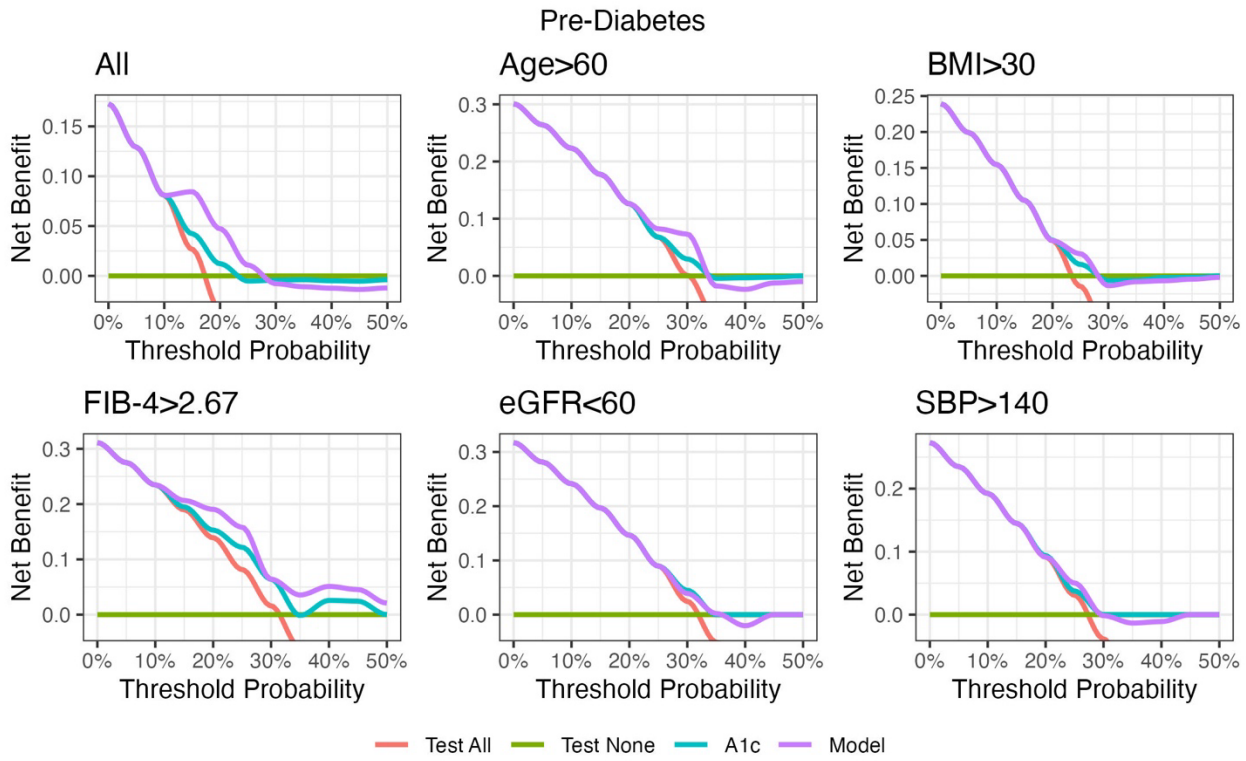
Supplementary Figure 5. Addition of the Full model with all variables does not provide significant difference from the minimal model for detection of pre-diabetes or diabetes.



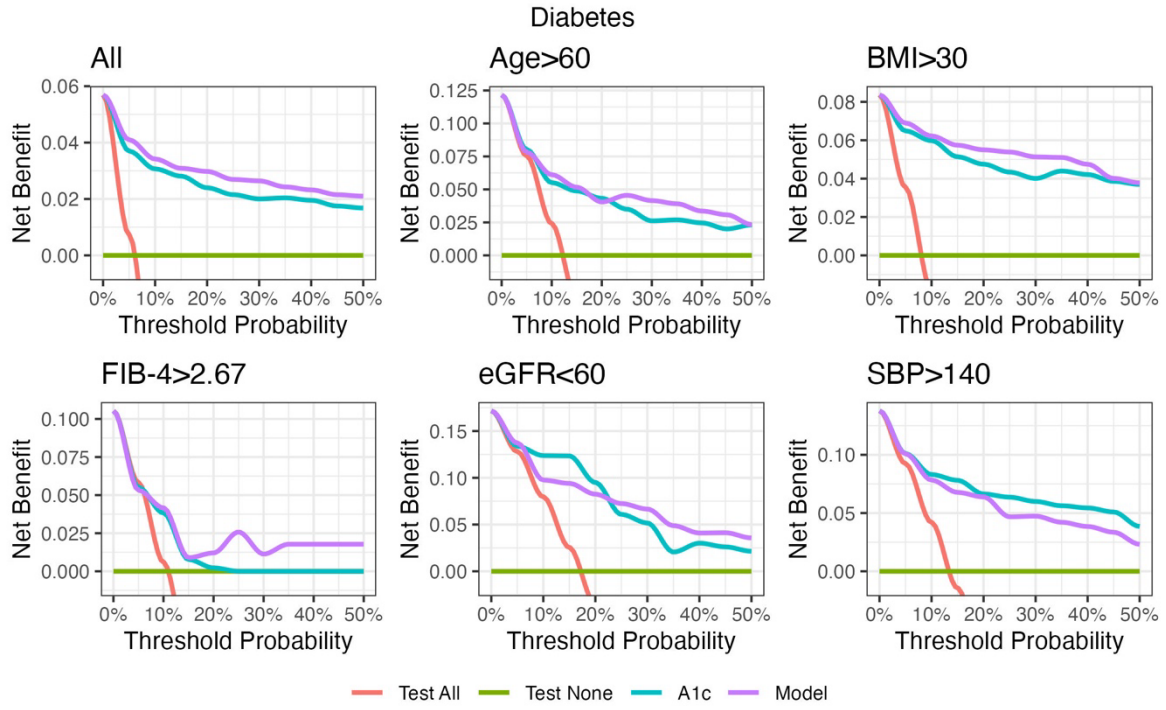
Supplementary Figure 8. Using the Training and Evaluation datasets, the Model outperforms the A1c for both diabetes and pre-diabetes diagnosis. Receive Operator Curves of the Model vs A1c (non-red vs red lines, respectively) for predicting two-hour glucose. Inset table shows the sensitivity, specificity, and ROC for diagnosis of pre-diabetes and diabetes based on use of the Model, A1, and the Training or Evaluation datasets.

Supplementary Table 7. AUROCs for different cohorts comparing the hemoglobin A1c, the 20-feature Model, and the Full Model (Full).

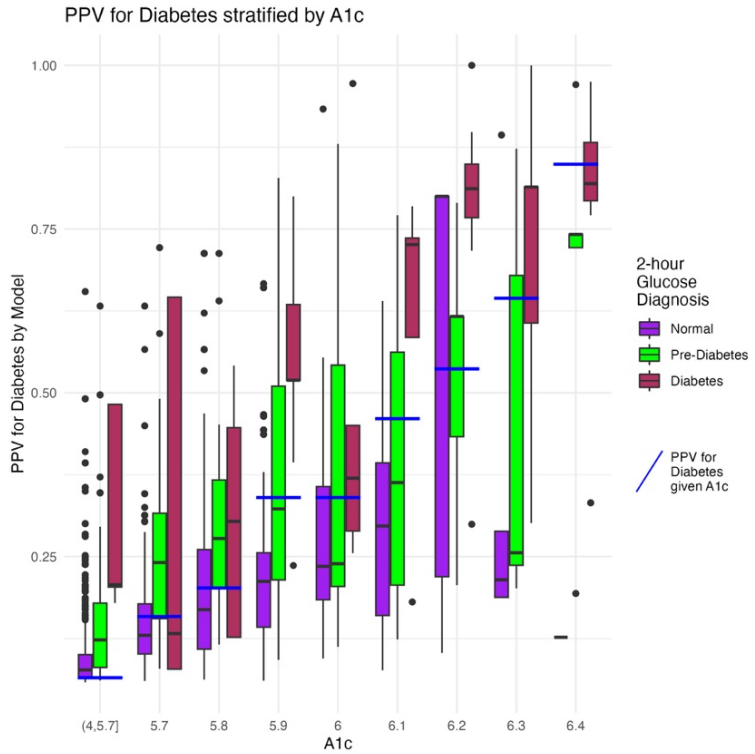
			AUROC - Pre-Diabetes				**AUROC - Diabetes**			
Cohort	**Training Size**	**Evaluation Size**	**A1c**	**FPG**	**Model**	**Full**	**A1c**	**FPG**	**Model**	**Full**
All	11718	2082	0.665	0.706	0.769	0.764	0.871	0.875	0.914	0.915
Age>60	2487	460	0.574	0.611	0.654	0.664	0.785	0.795	0.784	0.790
BMI>30	3481	694	0.632	0.671	0.683	0.686	0.925	0.93	0.955	0.956
FIB4>2.67	315	64	0.607	0.629	0.689	0.725	0.667	0.904	0.662	0.727
eGFR<60	496	91	0.536	0.615	0.586	0.562	0.834	0.588	0.830	0.849
SBP>140	1213	263	0.549	0.546	0.631	0.658	0.847	0.821	0.814	0.817



Supplementary Figure 6. Decision curve analysis for the diagnosis of pre-diabetes for the chronic disease cohorts.

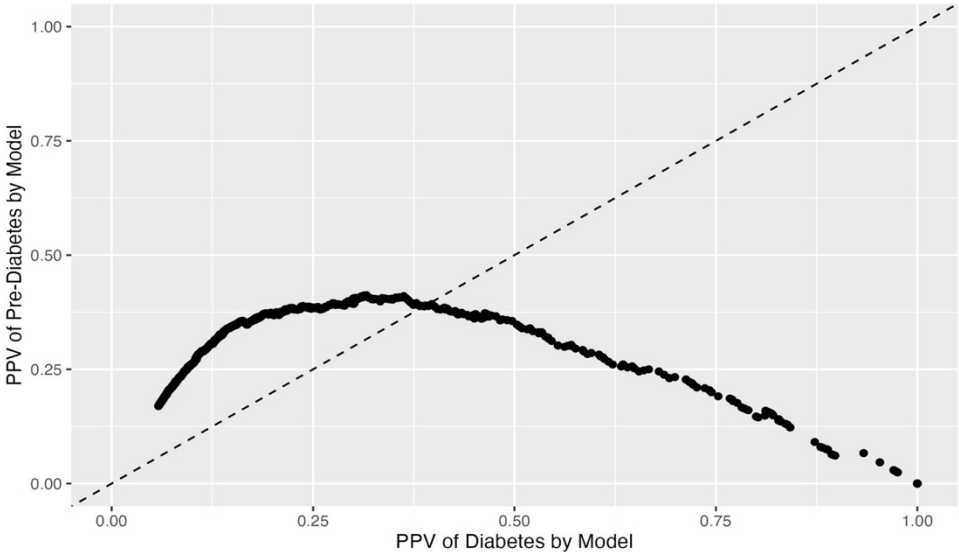


Supplementary Figure 7. Decision curve analysis for the diagnosis of diabetes for the chronic disease cohorts.

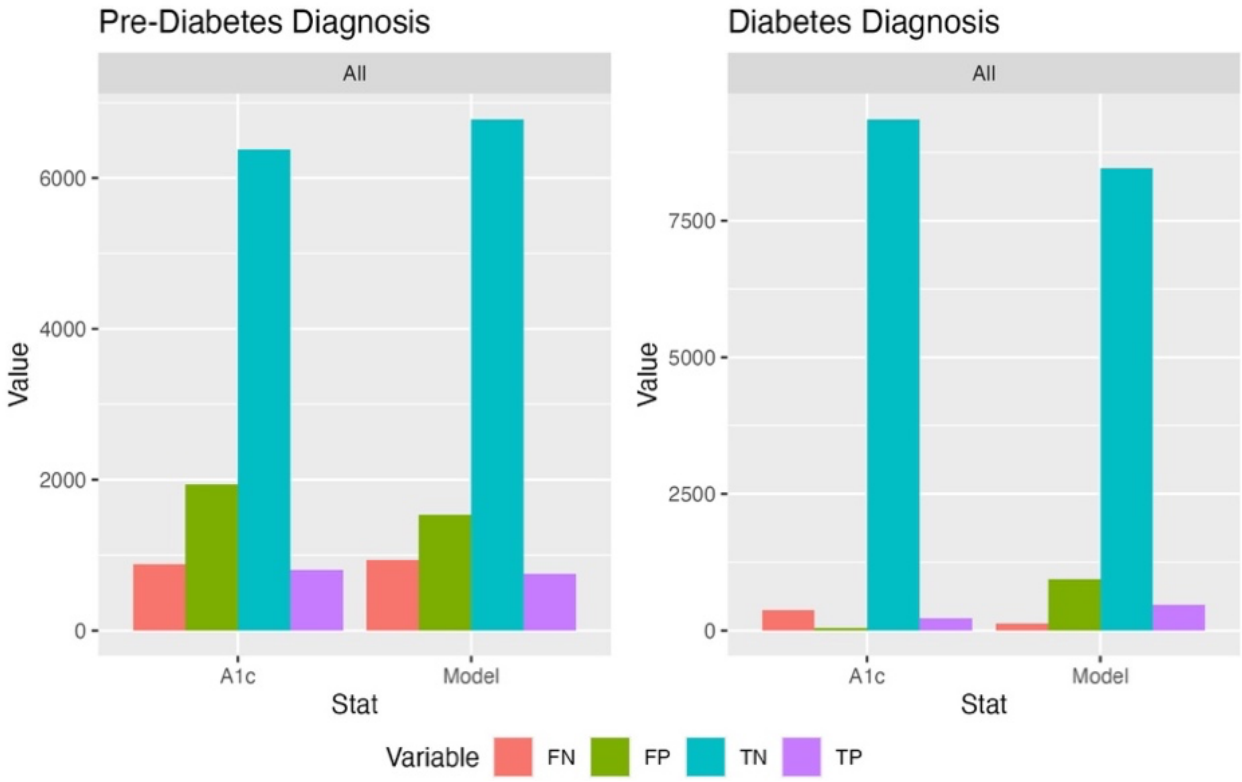


Supplemental Figure 8. Comparison of Positive Predictive Values (PPV) for A1c.

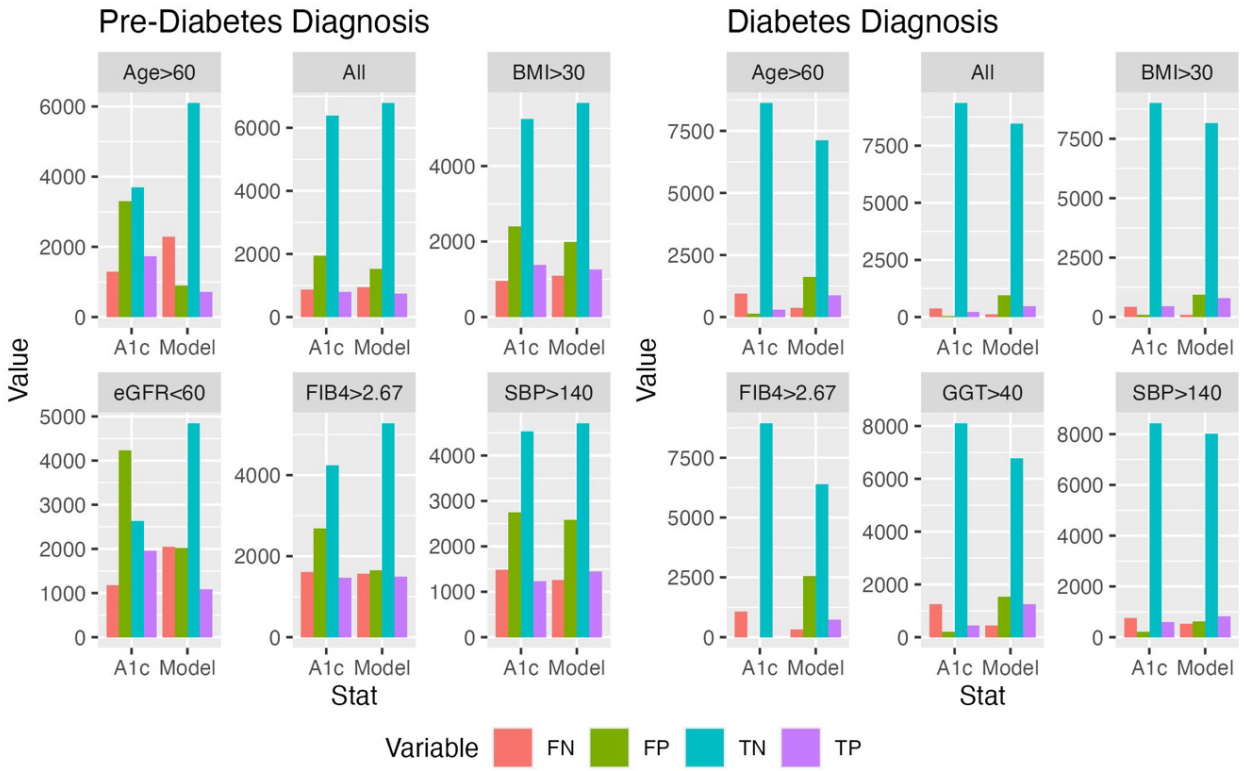
Supplemental Figure 9. Flow diagram of participants with A1c < 6.5% categorized using the Positive Predictive Value (PPV) from gold-standard oral glucose tolerance test (OGTT). The OGTT strata is marked as with the diagnosis "Normal", "Prediabetes (Pre-DM)", and "Diabetes (DM)" to the different models including A1c, where PPVs are grouped by <39%, 39-70%, and 100%. Flow colors indicate OGTT 2-hour glucose ≥ 200 mg/dL (purple), 140-200 mg/dL (green) or <140 mg/dL (blue). The OGTT strata are shown as the left-most and right-most columns to facilitate comparison with the Min model and the A1c model.



Supplementary Figure 10. For the Model, PPV of 39% there is a higher likelihood of a participant having Diabetes than having Pre-Diabetes.



Supplementary Figure 11 Comparison of False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) results for the 20-feature Model and the hemoglobin A1c, for the different chronic disease cohorts



Supplementary Figure 12. Comparison of False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) results for the 20-feature Model and the hemoglobin A1c, for the different chronic disease cohorts.

References

1. Diabetes NIo, Digestive, Diseases K. Diabetes Statistics. 2023; Available from: https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics
2. Parker ED, Lin J, Mahoney T, Ume N, Yang G, Gabbay RA et al. Economic Costs of Diabetes in the U.S. in 2022. Diabetes Care 2024;47(1):26-43.
3. Control CfD, Prevention\. Data & Research | Chronic Kidney Disease. 2024.
4. Diabetes NIo, Digestive, Diseases K. Overweight & Obesity Statistics. 2024; Available from: https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity
5. Acquavella J, Mehra R, Bron M, Suomi JM, Hess GP. Prevalence of narcolepsy and other sleep disorders and frequency of diagnostic tests from 2013-2016 in insured patients actively seeking care. J Clin Sleep Med 2020;16(8):1255-1263.
6. Hanas R, John G, International HBAcCC. 2010 consensus statement on the worldwide standardization of the hemoglobin A1C measurement. Diabetes Care 2010;33(8):1903-1904.
7. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D et al. 6. Glycemic Targets: Standards of Care in Diabetes-2023. Diabetes Care 2023;46(Suppl 1):S97-S110.
8. Collaboration NCDRF. Global variation in diabetes diagnosis and prevalence based on fasting glucose and hemoglobin A1c. Nat Med 2023;29(11):2885-2901.
9. Tseng E, Segal JB, Maruthur NM. Fasting Status of Patients Undergoing Ambulatory Laboratory Testing. Diabetes Care 2019;42(8):e133-e134.
10. Klein KR, Lingvay I, Tuttle KR, Flythe JE. Glycemic Management and Individualized Diabetes Care in Dialysis-Dependent Kidney Failure. Diabetes Care 2025;48(2):164-176.
11. Malkani S, Mordes JP. Implications of Using Hemoglobin A1C for Diagnosing Diabetes Mellitus. The American Journal of Medicine 2011;124(5):395-401.
12. Thomas MC. The clustering of Cardiovascular, Renal, Adipo-Metabolic Eye and Liver disease with type 2 diabetes. Metabolism 2022;128:154961.

13. Sharif A, Chakkera H, de Vries APJ, Eller K, Guthoff M, Haller MC et al. International consensus on post-transplantation diabetes mellitus. Nephrol Dial Transplant 2024;39(3):531-549.
14. Paul JC, Fitzpatrick JJ. Postpartum glucose screening among women with gestational diabetes. Appl Nurs Res 2020;56:151341.
15. Brown SD, Hedderson MM, Zhu Y, Tsai AL, Feng J, Quesenberry CP et al. Uptake of guideline-recommended postpartum diabetes screening among diverse women with gestational diabetes: associations with patient factors in an integrated health system in USA. BMJ Open Diabetes Res Care 2022;10(3).
16. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the Time Needed to Provide Adult Primary Care. J Gen Intern Med 2023;38(1):147-155.
17. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes Metab Res Rev 2000;16(3):164-171.
18. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009;151(11):775-783.
19. Chen T, Guestrin C. XGBoost. In. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.
20. Hu C, Li L, Huang W, Wu T, Xu Q, Liu J et al. Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. Infect Dis Ther 2022;11(3):1117-1132.
21. Zhang Q, Xu L, He W, Lai X, Huang X. Survival prediction for heart failure complicated by sepsis: based on machine learning methods. Front Med (Lausanne) 2024;11:1410702.
22. Zheng J, Li J, Zhang Z, Yu Y, Tan J, Liu Y et al. Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: a 10-year (2012-2021) Multicenter Retrospective Case-control study. BMC Gastroenterol 2023;23(1):310.
23. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.
24. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, Abdelmalek MF, Caldwell S, Barb D et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023;77(5):1797-1835.
25. He J, Lin Z, Song C, Zhang R, Wang H, Yuan S et al. High absolute neutrophil count with type 2 diabetes is associated with adverse outcome in patients with coronary artery disease: A large-scale cohort study. Front Endocrinol (Lausanne) 2023;14:1129633.
26. Rodriguez BAT, Johnson AD. Platelet Measurements and Type 2 Diabetes: Investigations in Two Population-Based Cohorts. Front Cardiovasc Med 2020;7:118.
27. Sabanayagam C, Shankar A, Li J, Pollard C, Ducatman A. Serum gamma-glutamyl transferase level and diabetes mellitus among US adults. Eur J Epidemiol 2009;24(7):369-373.
28. Haghighi S, Amini M, Pournaghshband Z, Amini P, Hovsepian S. Relationship between gamma-glutamyl transferase and glucose intolerance in first degree relatives of type 2 diabetics patients. J Res Med Sci 2011;16(2):123-129.
29. Hsu CC, Lee CH, Wahlqvist ML, Huang HL, Chang HY, Chen L et al. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. Diabetes Care 2012;35(11):2286-2292.
30. Yen FS, Wei JCC, Liu JS, Hwu CM, Hsu CC. Parental Income Level and Risk of Developing Type 2 Diabetes in Youth. JAMA Netw Open 2023;6(11):e2345812.
31. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003;26(3):725-731.
32. Hutchison AL, Cifu AS, Pillai AA. Discarding Dichotomization: retrieving Data in the service of Patient Care. Clinical Liver Disease 2022;19(1):12.

33. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 2019;140(11):e596-e646.
34. Jin J. Lipid Disorders: Screening and Treatment. JAMA 2016;316(19):2056-2056.
35. Reutrakul S, Zaidi N, Wroblewski K, Kay HH, Ismail M, Ehrmann DA et al. Sleep disturbances and their relationship to glucose tolerance in pregnancy. Diabetes Care 2011;34(11):2454-2457.
36. Tasali E, Van Cauter E, Hoffman L, Ehrmann DA. Impact of obstructive sleep apnea on insulin resistance and glucose tolerance in women with polycystic ovary syndrome. J Clin Endocrinol Metab 2008;93(10):3878-3884.
37. Rosenberg EA, Seely EW, James K, Arenas J, Callahan MJ, Cayford M et al. Relationship between carbohydrate intake and oral glucose tolerance test results among pregnant women. Diabetes Res Clin Pract 2021;176:108869.
38. Lee J, Kim MH, Jang JY, Oh CM. Assessment HOMA as a predictor for new onset diabetes mellitus and diabetic complications in non-diabetic adults: a KoGES prospective cohort study. Clin Diabetes Endocrinol 2023;9(1):7.

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	
	3c	D;E	Describe any known health inequalities between sociodemographic groups	
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	
	6b	D;E	Describe the eligibility criteria for study participants	
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	

[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

[2] Separately for all model building approaches.

[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	
Data sharing	18e	D;E	Provide details of the availability of the study data	
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

[4] This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.
[5] This relates to the code to implement the model to get estimates of risk for a new individual.