**Title Page:**

Diabetes e-alert System: A supervised machine learning system to predict patients with poor glycaemic control using routinely collected clinical data

David K Ryan MBChB (Hons) B Med Sci (Pharmacology)

Academic Foundation Doctor, St George’s University, London, UK.

[davidkdryan@gmail.com](mailto:davidkdryan@gmail.com)

August 2020

**Abstract**

**Objective:** Type 2 Diabetes Mellitus (T2DM) is often detected too late in its clinical course with many patients presenting with complications of unrecognised T2DM at the time of diagnosis. To address this, an e-alert system could be implemented within electronic healthcare record (EHR) systems to notify clinicians of patients with unrecognised or poorly managed T2DM. This study sought to develop a prototype e-alert system using supervised machine learning models in a large, community-based database to stimulate EHR data.

**Research Design and Methods:** Data from 23, 310 participants in the American National Health and Nutrition Examination Survey (2007 – 2016) were extracted to build binary classification models. This model was trained to classify participants based on measures of dysglycaemia (composite of either abnormal oral glucose tolerance test or elevated HbA1c ≥ 48 mmol/mol). Features were selected based on data typically available in an EHR including demographics, laboratory data, body measurements and prescribed medication. Five different models were evaluated using the F2 score through a 5-fold cross validation.

**Results**: A CatBoost model provided the optimal results (sensitivity 85.63%, specificity 90.65%, precision 59.31%, AUC-ROC 0.96). This model included: random blood glucose, age, LDL cholesterol, metformin, gamma glutamyl transferase, BMI, triglycerides, chloride, ethnicity, insulin, creatinine and bicarbonate. Unrecognised T2DM was prevalent in this cohort with over 40% of participants with biochemical evidence of T2DM not acknowledging a diagnosis of diabetes. The system made 462 alerts when applied to the test set (n=2,331), detecting 67.82% of the participants with unrecognised T2DM. The system missed 14% of opportunities when patient care could have been improved but 60% of all alerts would be clinically beneficial.

**Conclusions**: It is possible to identify patients with poor glycaemic control to a high degree of sensitivity and specificity using routinely collected clinical data. However, further testing with real-life EHR data and ultimately, clinical trials would be required prior to implementation of such a system.

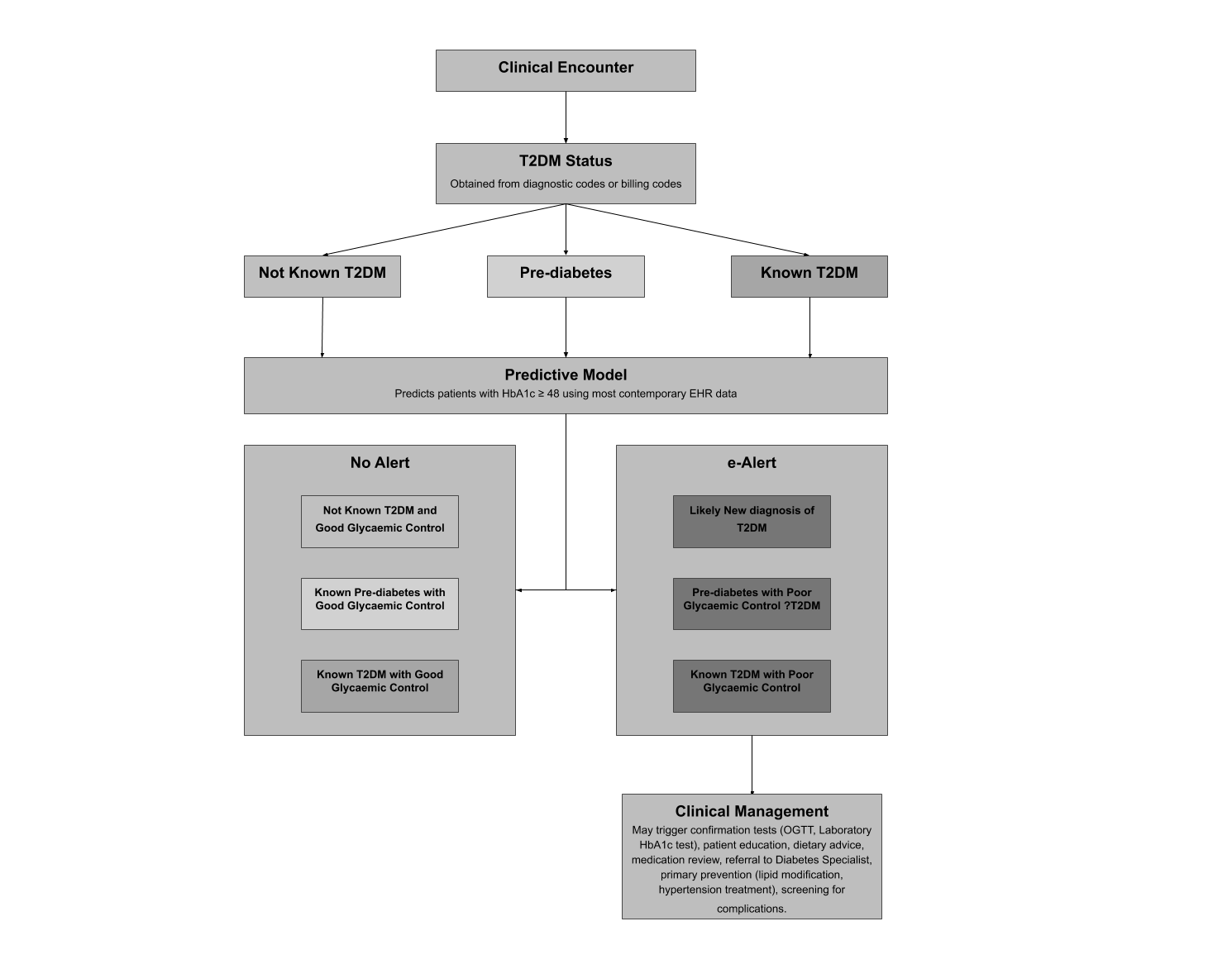
Type 2 Diabetes Mellitus (T2DM) affects approximately 1 in 10 adults in the UK,1 with 422 million people affected worldwide.2 Type 2 Diabetes Mellitus (T2DM) is largely an insidious disease process that can remain undetected for many years prior to diagnosis.3,4 Studies involving patients with diabetic retinopathy suggest that the effects of dysglycaemia can start 9 – 12 years prior to disease recognition, resulting in a significant, but largely avoidable burden of disease at time of diagnosis.5 Therefore, it is unsurprising that over 20% of people with T2DM are unaware of their condition.6 Missed or late diagnosis of diabetes represents a significant burden on patients and on wider healthcare systems.3 Given the large numbers of people at risk of T2DM and the growing evidence for the reversibility of T2DM with intensive dietary and lifestyle interventions,7 the timely identification of T2DM is now a key priority.

With the increasing use of electronic healthcare records (EHR), there is now a growing potential to use machine-learning algorithms to improve real-time detection of chronic diseases and promote safer patient care.8–10 Digital transformation of healthcare systems through predictive models and decision support systems has been identified as a key priority within the UK National Health Service’s long-term plan.11 Implementing such technology within EHRs has already shown promise in better identification of acute kidney injury (AKI) and reduced mortality rates associated with hospital-acquired AKI.12 In other settings, EHR based triggers are used to notify clinicians of missed venous thromboembolism prophylaxis,13 clinical deterioration,14 and prescribing errors.15,16 EHRs are no longer inert repositories of medical information, but are beginning to play a key role in promoting safer and more efficient healthcare and provide a rich data source for knowledge-discovery.16,17

Applied to the problem of earlier recognition and better management of T2DM, there is a potential to integrate safety systems within an EHR architecture to predict patients who may have T2DM, identify patients who may require better diabetic management or have complications of T2DM. This system would collect contemporary data from the EHR and at each interaction it would determine whether the patient in question would be predicted to have poor glycaemic control – alerting the clinician to consider formal T2DM screening, education, treatment or onwards referral to a Diabetes Specialist (figure 1).

Several machine-learning classification algorithms such as support vector machines,18,19 decision trees,20 rotation forest models,21 random forest classifiers and gradient-boosted decision trees22 have shown promise in identifying cases of T2DM with high levels of accuracy.17 However many of these models are limited in that they use specialist and oftentimes unrealistic clinical features that would not be reliably recorded in a typical clinical encounter. One recent study developed a classifier for identifying patients with T2DM which achieved a recall rate and precision of 89%.22 However, this study included features such as dietary sodium, fibre, carbohydrate, calcium and caffeine intake.22 Other models include unrealistic physical examination metrics (triceps skin fold thickness),18 questionnaire data that may not be readily available (measurements of physical activity and alcohol consumption),19 specialised laboratory tests (2 hour serum insulin)18 or require significant human feature-engineering.23 It is not clear how well these models would work if integrated into a live EHR as such data can be plagued with limited feature availability, poorly recorded subjective features (such as physical activity history), infrequent longitudinal data, and high-levels of incomplete, inconsistent or inaccurate information.24 Furthermore, previous studies have not considered how such a model would be incorporated within usual clinical care.

The aim of the present study is to develop a prototype diabetes e-alert system which could be implemented within an EHR architecture to improve the recognition of poorly controlled or unrecognised T2DM.

**Figure 1:** Proposed Diabetes e-alert System

**Figure 1:** Schematic of a diabetes e-alert system integrated within an electronic healthcare record (EHR) system. During each patient encounter, the system is harvesting relevant data from the EHR and classifying the patient according to whether they are predicted to have either good or bad glycaemic control. Dysglycaemia is defined as either an abnormal HbA1c or oral glucose tolerance test result. It takes account of whether the patient is known to have diabetes and adjusts the message the clinician receives based on this. The e-alert prompts the clinician to consider diagnostic testing, patient education, medication review, management of type 2 diabetes mellitus (T2DM) complications or onwards referral to a diabetes specialist.

**Research Design and Methods:**

**Data Source and Study Population:**

Data for the present study were obtained from the publicly available American National Health and Nutrition Examination Survey (NHANES) conducted annually by the United States National Center for Health Statistics since 1999. This is a cross-sectional survey that captures the health and nutritional status of a representative sample of the civilian, non-institutionalised US population. The NHANES study has four distinct components including questionnaires about demographics and health status, physical examination, laboratory tests and food diaries. However, due to changes in variables measured and coding of responses across different NHANES periods, our study was restricted to the past six NHANES surveys. This covered a time-period of 2007 – 2016 and a total population of 50, 588. Data was downloaded as SAS files from wwwn.cdc.gov/nhanes/Deafult.aspx.

Exclusion criteria for the present study included: age below 20 years, diagnosed with diabetes before 20 years, pregnancy (either self-declared or positive pregnancy urine test). These participants were excluded to avoid including participants who may have other forms of diabetes such as Type 1 Diabetes or Gestational Diabetes. Participants with no available laboratory data were also removed as it is unlikely that participants with no laboratory data would be included in a typical EHR. On average, approximately 10% of all features were missing. These values were imputed with an out-of-sample value (-999). Participants with no recorded oral glucose tolerance test or HbA1c measurement were also removed as this was the target variable. Imputation was not appropriate in this instance as it may represent a possible data leakage between train and test datasets.

**Target Definition**

The binary classification model was trained to detect participants with biochemical evidence of a dysglycaemia indicative of either a new diagnosis of T2DM or poorly managed T2DM. This was a composite of either an elevated HbA1c ≥ 48 mmol/mol or an abnormal oral glucose tolerance test (fasting value > 7.0 mmol/L or 2-hour result > 11.0 mmol/L). A positive case was defined as having either an abnormal HbA1c or oral glucose tolerance test. HbA1c value was obtained from variable reference LBXHGB and oral glucose tolerance test were extracted from variables LBDGLTSI and LBDGLUSI. World Health Organisation cut-off thresholds for diagnostic diabetes tests were used in the present study.25 Furthermore, a HbA1c value ≥ 48 mmol/mol is also a target used by the UK’s National Institute for Health and Care Excellence T2DM guidelines for people managed with lifestyle or diet modification and a single drug not associated with hypoglycaemia (such as metformin). 26

**Table 1**: Features extracted from National Nutrition and Health Examination Survey

|  |  |  |
| --- | --- | --- |
| **NHANES Code** | **Variable Details** |  |
| **SEQN** | Anonymous identification number |  |
| **RIAGEYR** | Age |  |
| **RIAGENDR** | Gender |  |
| **RIDRETH1** | Ethnicity |  |
| **BPXSY1, BPXDI1,**  **BPXSY2, BPXDI2,**  **BPXSY3, BPXDI3,**  **BPXSY4, BPXDI4** | Mean arterial pressure  (0.33\*average systolic pressure in mmHg + 0.67\*average diastolic blood pressure) |  |
| **BMXBMI** | BMI (kg/m2) |  |
| **LBXSATSI** | Alanine Aminotransferase (ALT) (U/L) |  |
| **LBDSALSI** | Albumin, refrigerated serum (g/L) |  |
| **LBXSAPSI** | Alkaline Phosphatase (ALP) (U/L) |  |
| **LBXSASSI** | Aspartate Aminotransferase (AST) (U/L) |  |
| **LBXSBU** | Blood Urea Nitrogen (mg/dL) |  |
| **LBXSCLSI** | Chloride (mmol/L) |  |
| **LBXSCH** | Cholesterol, refrigerated serum (mg/dL) |  |
| **LBDSCRSI** | Creatinine refrigerated serum (μmol/L). This was used along with age, gender and ethnicity to calculate estimated glomerular filtration rate according to the modification of diet in renal disease (MDRD) formula. |  |
| **LBXSGTSI** | Gamma Glutamyl Transferase (GGT) (U/L) |  |
| **LBDSGLSI** | Glucose, refrigerated serum (mmol/L) |  |
| **LBXSKSI** | Potassium (mmol/L) |  |
| **LBXSNASI** | Sodium (mmol/L) |  |
| **LBDSTBSI** | Total Bilirubin (μmol/L) |  |
| **LBDSCASI** | Total Calcium (mmol/L) |  |
| **LBDSTRSI** | Triglycerides, refrigerated serum (mg/dL) |  |
| **LBDLDL** | LDL cholesterol (mmol/L) |  |
| **LBDHDDSI** | HDL cholesterol (mmol/L) |  |
| **RXDRUG** | Prescription list searched for insulin, metformin, other oral anti-diabetic agents (sitagliptin, liraglutide, gliclazide, gilclazide, glipizide, tolbutamide, gilbenclamide), insulin (corroborated with variable DIQ050 – taking insulin now), statin, ACE-inhibitor/angiotensin II receptor blocker. |  |
| **LBXGHB** | Glycated Hb (%) adjusted to mmol/mol. Participants with a haemoglobin in the anaemic range (LBXHGB - male Hb < 140 mg/L or female Hb < 120 mg/L) were imputed as having missing value for HbA1c. |  |

Table 1 shows data extracted from the NHANES study according to NHANES code and any subsequent modification to variables. Data converted from SAS files available at: wwwn.cdc.gov/nhanes/Deafult.aspx

**Machine Learning Models**

In this study, several tree-based, supervised machine learning algorithms will be used. These include decision trees, random forest, XGBoost (gradient boost ensemble model), CatBoost and the Amazon autogluon autoML model for tabular prediction. These models are suitable for exploring non-linear relationships and are not overtly affected by multi-collinearity.

Decision Trees

Decision trees refer to model that classify based on separating cases (or in this context, survey participants) at various nodes according to the values of different features.27 The model seeks to maximise information gain at each level of the tree, splitting cases to minimise entropy and create clear distinctions between cases.27

Random Forest

Random forest are based on the principle of combining multiple weak base decision trees in an ensemble with sub-selection of features within each base tree to reduce model over-fitting.22

Extreme Gradient Boosting (XGBoost)

This model improves on random forest models through a boosting process, where individual decision trees are sequentially optimised based on errors in previous decision trees.22 This uses a gradient descent algorithm to minimise the loss function associated with addition of new trees to the ensemble model.22

CatBoost Model

CatBoost is a new, open-sourced gradient boosting model that has shown state-of-the-art results on various classification problems.28 Boosting processes introduce an inherent overfitting into ensemble models due to prediction shift.28 This occurs as the base models are continuously being optimised through gradient descent on the same set of training variables, representing a form of data leakage.28 Furthermore, tree-based models oftentimes re-code categorical data into target statistics, which can contain information such as the mean of the target variable and are also a form of target leakage.28 To avoid these issues, CatBoost uses ordered boosting where it divides data set into random permutations and uses these random sets to train and optimise the base models of the ensemble.28 For categorical features, it also creates random permutations of data for creating target statistic and training models, again reducing target leakage.28

Amazon AutoML Autogluon Model for Tabular Data Prediction

Traditional machine learning is a labour intensive process and requires human input for feature engineering and model selection.29 There is now growing interest in speeding up the process of machine learning and making it more accessible to domain experts using automated machine learning (autoML). AutoML aims to automate the pipeline for machine learning problems and reduce the human involvement in model development.29 AutoML iterates through several different approaches and pre-existing models to determine optimal feature engineering, model selection and parameters for the problem in question. Amazon’s version of autoML, Autogluon, automatically recognises problem-types and fits various models (KNN, random forest, lightGBM, XGBoost, CatBoost, neural networks and ensemble models among others).30

**Data Organisation**

5% of the available datasets were set aside as a validate set and 10% of total dataset was removed for final test set (figure 2). This separation is important to avoid data leakage and making any decision in the modelling process using the test set. Data splitting was stratified to ensure that the train, validate and test datasets have the same percentage of positive and negative cases. Data was scaled to account for different measurement units using the mean and standard deviation of the training samples alone to avoid data leakage. Categorical features were one-hot encoded for all models apart from CatBoost, which has an inbuilt one-hot encoder within the model architecture. No data processing was done for the Amazon autoML modelling as this has intrinsic strategies to encode and process data.

**Feature Selection**

Features were rationally chosen based on what would be reliably recorded within the EHR for most routine healthcare interactions (table 1). This includes demographic information, routine laboratory tests, measurements such as blood pressure and BMI and prescribed medication. Such variables would also be readily available in different EHR formats – aiding interoperability of the model.16

Models (decision trees, random forest, XGBoost, CatBoost models) were initially built using all the available features. To determine the optimal number of features in each model, we developed a forward-selection process where, for each model, we ranked each feature in order of decreasing SHAP importance value. Sequential models were constructed, starting with a model consisting of the most important feature and adding further features to the model in order of decreasing importance. We then evaluated the impact of adding further feature to the model using the F2 statistic. A graph was constructed to show the gain in F2 statistic for each feature added to the model, and the optimal number of features were selected based on where the F2 score plateaued. Each model was then trained using the optimal sequence of features. No data processing was used in the Amazon AutoML implementation as this has intrinsic methods to manage feature engineering, data pre-processing, model selection and parameter tuning.

**Model development**

Model parameters were fine-tuned using randomised search of unique parameters for each model (iterations = 250, cross-validation process = 2, see coding documentation for further details). Final models were scored and selected on basis of highest F2 score. The F2 score is a variation on the F1 metric but places a higher weight on sensitivity over precision. This is particularly appropriate for the current classification problem where sensitivity, (or the avoidance of false negatives) is more clinically important than precision of diagnosis.

Standard classification systems use a threshold of 0.5 to distinguish between binary classes, however due to class-imbalance in the present dataset, this threshold was adjusted to reduce false negative rates and promote sensitivity within the model. The probability threshold was established using the validation dataset, calculating F2 score for all probability thresholds between 0.01 and 0.99, in increments of 0.01 and selecting the probability threshold which resulted in the best F2 score. This threshold was selecting using the validate set to avoid any possible leak of testing and training data and to avoid making decisions about the model based on test data. This threshold was then applied to the model for the train dataset for evaluation of final model performance. All model predictions for validate and test sets were made with a cross-validation process (cross-validation=5).

Models were evaluated using a standard battery of tests including sensitivity, specificity, positive predictive value, accuracy, area under the receiver operator curve (table 2a). Confusion matrices were made for all models based on outcomes in the hold-out test set (table 2b).

**Figure 2:** Work Flow

Total Cohort NHANES 2007 – 2016   
(n = 50, 588)

Data pre-processing:

Participants excluded:

Pregnant (n = 317)

Age < 20 years (n = 21, 869)

Diabetes diagnosis < 20 years old (n = 129)

Containing no laboratory measurement (n = 2, 547)

No HbA1c or oral glucose tolerance test recorded (n = 2, 416)

Excluded participants   
(n = 27, 278)

Eligible participants   
(n =23, 310)

)

Data processing:   
Variables calculated, missing data imputed with out-of-sample values

Test set held (10%)

(n = 2, 331)

Data split into train (85%), validate (5%) and test (10%) sets and standard scale applied

Model baseline scoring and randomised search for parameter optimisation on train data (n = 19, 814)

Probability threshold for classification adjusted to maximise F2 score on validate set (n = 1, 165)

Evaluation metrics on the hold-out test set (n = 2, 331)

**Figure 2:** Schematic showing the workflow for obtaining and cleaning data as well as dividing the data into train, validate and test sets.

**Table 2a:** Evaluation Metrics

|  |  |
| --- | --- |
| **Metric** | **Formula** |
| **Sensitivity (recall)** | TP / (TP + FN) |
| **Specificity** | TN / (TN + FP) |
| **Positive Predictive Value (precision)** | TP / (TP + FP) |
| **Accuracy** | (TP + TN)/(TP + TN + FP + FN) |
| **Area under the receiver operator curve** | Evaluation metric that summarises performance of a model at all classification thresholds. Receiver operator curve is established by plotting true positive rate on Y axis with false positive rate on the x axis. A model with an AUC of 1 would classify all cases correctly. |
| **Fβ-score** | Summary statistics for a model which places a greater weight on recall over precision.  Fβ = (1 + β2) x (precision x recall)  (β2 x precision) + recall |

Table 2a: Formulae for the evaluation metrics used in the present study. TN: true negative, TP: true positive, FN: false negative, FP: false positive. Fβ is the generalised formula and for F2 score β = 2.

**Table 2b:** Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | **Predicted A** | **Predicted B** |
| **Actual A** | True positive | False positive |
| **Actual B** | False negative | True negative |

Table 2b: Confusion matrix shows how the final model classifies participants in the hold-out test set. Green shading denotes correct classification (true positive and true negative cases), amber shading denotes false positives and red shading refers to false negatives. In this context, a false negative is the most serious misclassification as it would result in a patient with poor glycaemic control not being detected by the model. An amber classification (false positive) has less patient safety concerns but represents a burden for healthcare systems who have to delineate which of the positive cases are in fact true.

**Feature Importance**

To explore the output of the models, SHAP (Shapley additive explanations) values were employed as a means of assessing feature importance. SHAP values are calculated by comparing models where a feature is included and a similar model where the feature is omitted.31 This is repeated for all possible combinations of features within the model and aggregated to calculate an overall feature importance.31 The benefit of SHAP values are that they take into account interactions between multiple variables.31 SHAP summary plots were generated to depict what feature value led to a certain classification.

**Exploratory Data Analysis**

Features were further explored by comparing median values for both participants with and without T2DM. Statistical differences between cohorts were calculated using the Chi-square test for categorical data and Mann-Whitney test for continuous variables (significance set at p < 0.05)

**Ethical Considerations**

The NHANES survey is a publicly available dataset and is approved by the National Center for Health Statistics Research Ethics Review Board. All study participants provided written informed consent. No institutional approval was required for the present study.

**Coding Files and Data Availability**

All models were implemented using python version 3.7.4 with scikit-learn library version 0.20.1 (decision tree, random forest classifier, random search cross validation), xgboost classifier version 0.90 and CatBoost version 0.22 for machine learning algorithms. Amazon autogluon AutoML is commercially available. Coding files and data available at: <https://github.com/dkryan/diabetes_ml>.

**Results**

**Demographics**

The final dataset consisted of 23, 310 participants, of which 2, 747 (11.8%) stated that they had a prior diagnosis of diabetes. A further 514 participants stated that they had pre-diabetes (2.20%). Table 3 describes the demographics of the dataset, segregated according to self-identified diabetic status. Participants with previously diagnosed T2DM represented an older cohort, with a higher proportion of participants from black and ethnic minority groups.

**Diabetic Status**

In total, 3, 197 (13.72%) patients had biochemical evidence of T2DM, with 2, 033 participants (8.72%) having an elevated HbA1c, 2, 022 (8.67%) participants having an abnormal oral glucose tolerance test (OGTT). There were 858 participants who had evidence of T2DM on both HbA1c and OGTT tests. Average HbA1c levels for participants who stated they had a prior diagnosis of T2DM was 58.88 mmol/mol (standard deviation 20.50 mmol/mol), compared to a value of 36.67 mmol/mol (standard deviation 6.71 mmol/mol) for participants who stated they did not have a known diagnosis of diabetes and 43.40 mmol/mol (standard deviation 10.33 mmol/mol) for participants who said they had pre-diabetes.

There was a high rate of unrecognised diabetes in this cohort with over 40% of participants (n = 1, 316 of 3, 197) with biochemical evidence of T2DM not acknowledging a diagnosis of diabetes. The mean HbA1c value for participants with unrecognised T2DM was 51.01 mmol/mol (standard deviation 17.27), compared to a value of 66.19 mmol/mol (standard deviation of 20.08) for participants with recognised T2DM. Mean value for participants without biochemical evidence of T2DM was 36.11 mmol/mol (standard deviation 4.31). Of these participants with unrecognised diabetes, only 152 (11.56%) of this cohort had an isolated elevated random blood sugar level recorded (random blood glucose > 11.1 mmol/L).

Participants with unrecognised T2DM were prescribed statin therapy at significantly lower rates compared to participants whose T2DM was known (27.96% vs. 52.65%, p<0.001). Similarly, ACE-inhibitors and angiotensin II receptor blockers were prescribed at much lower rates in patients with unrecognised T2DM compared to patients who were aware of their diabetic status (27.96% vs. 54.28%, p < 0.001).

**Table 3:** Demographic Profile of Dataset

There are significant differences in terms of the demographic features of participants with and without T2DM (table 3).

|  |  |  |  |
| --- | --- | --- | --- |
| **Demographic Factor** | **Participant without known T2DM  (n = 20, 049)** | **Participant with known T2DM**  **(n = 2, 747)** | **P value** |
| **Gender**       Male       Female | 9, 543 (47.60%)  10, 506 (52.40%) | 1, 326 (48.27%)  1, 421 (51.73%) | p = 0.508 |
| **Ethnicity**       Non-Hispanic White       Non-Hispanic Black       Mexican American       Other Hispanic       Other race including multi-racial | 9, 023 (45.0%)  3, 492 (17.42%)  3, 090 (15.41%)  2, 183 (10.89%)  2, 261 (11.28%) | 974 (35.46%)  664 (24.17%)  516 (18.78%)  329 (11.98%)  264 (9.61%) | p < 0.001 |
| **Age (years)** | 46.0 (28.0) | 62.0 (18.0) | p < 0.001 |
| **First Degree Family History of T2DM**       Yes       No       Unknown/not recorded | 7, 189 (35.86%)  12, 476 (62.23%)  384 (1.92%) | 1, 836 (66.84%)  843 (30.69%)  68 (2.48%) | p < 0.001 |

**Table 3:** Demographic profile of participants in the dataset, grouped according to self-identified diabetic status. This table shows that participants who self-identified as having a diagnosis of diabetes tend to be older and from Black and Ethnic minority groups. Data presented as either value count (as a percentage of patients in either cohort) or median with interquartile range. Significance of differences for categorical data (gender, ethnicity and family history) between groups compared using chi-square statistic. Significance for continuous data (age) compared using Mann-Whitney test. Significance set at p < 0.05.

**Current Anti-Diabetic Treatment**

Over 25% of participants with a known diagnosis of T2DM were prescribed insulin (table 4). Of note, 10 patients who did not acknowledge a diagnosis of T2DM in the questionnaire segment of the NHANES study were prescribed insulin. Metformin was prescribed to 126 participants who did not have a diagnosis of T2DM.

**Table 4:** Treatment profile for participants with Known T2DM

|  |  |
| --- | --- |
| **Treatment Modality** | **Participants with known T2DM (n = 2, 747)** |
| Metformin  Other Anti-diabetic Agent ±  Insulin | 1, 750 (63.70%)  651 (23.70%)  731 (26.61%) |

**Table 4:** Treatment profile for participants with T2DM. Data presented as value count (and percentage of participants receiving treatment modality in the cohort with T2DM). ± Other anti-diabetic agent refers to sitagliptin, liraglutide, gliclazide, gilclazide, glipizide, tolbutamide, gilbenclamide. No treatment refers to not being prescribed at the time of study metformin, other anti-diabetic agent or insulin. This table does not account for previous treatment, intolerances of combination therapies.

A significantly higher proportion of participants with known T2DM were prescribed statins compared to participants without known T2DM (non-diabetic percentage 13.41% vs. diabetic percentage 53.11%, p < 0.001). Rates of ACE-inhibitors/angiotensin II receptor blockers were also higher in participants with T2DM compared to participants without T2DM (non-diabetic percentage 13.09% vs. diabetic percentage 53.22%, p < 0.001).

**Feature Selection**

A forward selection process was developed to determine the optimal features to be included in each respective model. The final models were trained using the features listed in table 5. For the CatBoost model, the maximum F2 value occurred at 12 features and from this point we can see that the addition of further features does not overly improve the model (figure 3).

**Figure 3:** Forward Selection of Features in the CatBoost Model

A screenshot of a cell phone

Description automatically generated

**Figure 3** shows the effect of adding further features to the CatBoost model. This model was classifying participants based on evidence of dysglycaemia (abnormal HbA1c or oral glucose tolerance test). Features were first ranked in order of importance according to a CatBoost model with all features included. Sequential models were then built starting with the most important feature and adding the next most important feature in the next model. The X axis shows the number of features included and the Y axis denotes the F2 value for the model in question. This shows the best performing model occurring with the top 12 most important features. These 12 features were: Glucose, refrigerated serum (mmol/L), age, LDL cholesterol (mmol/L), metformin, gamma glutamyl transferase (GGT), BMI, triglycerides, refrigerated serum (mg/dL), chloride (mmol/L), ethnicity, insulin. creatinine refrigerated serum (µmol/L), bicarbonate (mmol/L).

**Table 5:** Features included in each model

|  |  |
| --- | --- |
| **Decision Tree** | Glucose, refrigerated serum (mmol/L) |
| **Random Forest** | Glucose, refrigerated serum (mmol/L), age, metformin prescribed, metformin not prescribed, LDL cholesterol (mmol/L), gamma glutamyl transferase (GGT), ACE inhibitor prescribed, ACE inhibitor not prescribed, BMI, statin prescribed, insulin prescribed, triglycerides, refrigerated serum (mg/dL), HDL cholesterol (mmol/L) |
| **XGBoost** | Glucose, refrigerated serum (mmol/L), age, LDL cholesterol (mmol/L), BMI, metformin not prescribed, chloride (mmol/L), eGFR, mean arterial pressure, triglycerides, refrigerated serum (mg/dL), HDL cholesterol (mmol/L), alanine aminotransferase (ALT), gamma glutamyl transferase (GGT), cholesterol, refrigerated serum (mg/dL), alkaline phosphatase (ALP), creatinine, refrigerated serum (µmol/L). |
| **CatBoost** | Glucose refrigerated serum (mmol/L), age, LDL cholesterol (mmol/L), metformin, gamma glutamyl transferase (GGT), BMI, triglycerides, refrigerated serum (mg/dL), chloride (mmol/L), ethnicity, insulin. creatinine refrigerated serum (µmol/L), bicarbonate (mmol/L). |

**Table 5:** Table showing optimised features that were used to train each model. Features were selected using a forward stepwise approach. Each model was initially built with all features included and following this, features were ranked in order of decreasing importance. Sequential models were then developed where features were added, starting with the most important feature and ending with a model including all features. The effect of adding further feature to the model was tested using the F2 statistic and the optimal number of features was selected based on the model that provided the best F2 statistic. Note no feature selection was employed for Amazon AutoML as this implementation has inherent methods for automating data processing, model selection and optimisation. For decision tree, random forest and XGBoost models, categorical features were one-hot encoded which explains why the presence or absence of a prescription medication (e.g. metformin, ACE inhibitor/Angiotensin Receptor Blocker, statin, other oral anti-diabetic agents) are included separately in the model.

**Model Evaluation:**

Based on the F2 score, a CatBoost Model represented the most effective model to classify whether a participant had evidence of dysglycaemia (HbA1c ≥ 48 mmol/mol/L or an abnormal oral glucose tolerance test - either a fasting value > 7.0 mmol/L or 2-hour result > 11.0 mmol/L). This model had a sensitivity of 86.10% and specificity of 95.74%) when applied to the test set (n = 2, 331). The positive predictive value for this model was 59.31% (table 6).

Of the hold-out test set (n = 2, 331), the model misclassified 46 participants as false negatives (table 7).

**Table 6:** Evaluation metrics for models developed using all available features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **ROC\_AUC** | **Sensitivity** | **Specificity** | **Positive Predictive Value** | **F2 Score** |
| Decision Tree | 0.89 | 0.7813 | 0.9075 | 0.5734 | 0.7284 |
| Random Forest | 0.95 | 0.8783 | 0.8727 | 0.5233 | 0.7733 |
| XGBoost | 0.93 | 0.8031 | 0.9194 | 0.6134 | 0.7563 |
| **CatBoost** | **0.96** | **0.8563** | **0.9065** | **0.5931** | **0.7865** |
| AutoML  (weighted ensemble model) | 0.96 | 0.8820 | 0.8947 | 0.5182 | 0.7734 |

Table 6 shows the evaluation metrics for models developed using all available features. The optimal model was selected based on the highest F2 score (random forest F2 0.7982). ROC\_AUC: area under the receiver-operator curve.

**Table 7:** Confusion Matrix for CatBoost Model applied to the hold-out test set (n = 2, 331)

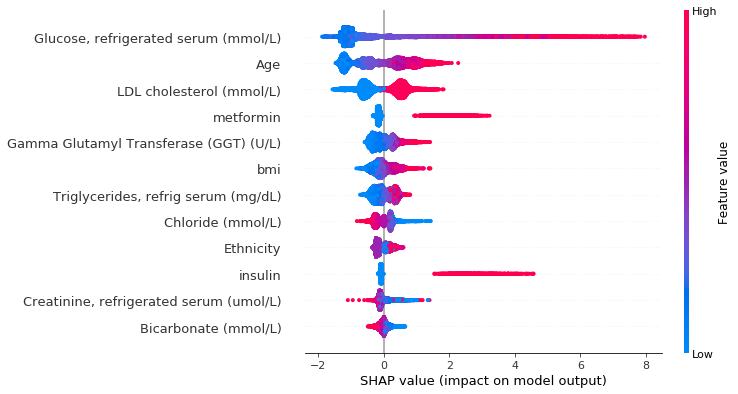
|  |  |  |
| --- | --- | --- |
|  | **Predicted good glycaemic control** | **Predicted poor glycaemic control** |
| **Actual good glycaemic control** | 1823 | 188 |
| **Actual poor glycaemic control** | 46 | 274 |

Table 7 is a confusion matrix showing the classification of participants in the hold-out test set (n = 2, 331). Poor glycaemic control is a composite outcome comprising of either an abnormal HbA1c value (HbA1c ≥ 48mmol/mol) or an abnormal oral glucose tolerance test (either a fasting value > 7.0 mmol/L or 2-hour result > 11.0 mmol/L). True positive: 274, true negative: 1, 823, false positive: 188, false negative: 46.

**Feature Importance:**

SHAP feature importance values for the model shows that the features influencing classification were, in order of decreasing importance: random blood glucose, age, LDL cholesterol, metformin, gamma glutamyl transferase (GGT), BMI, triglycerides (mg/dL), chloride (mmol/L), ethnicity, insulin, creatinine (µmol/L) and bicarbonate (mmol/L). High values for glucose, age, LDL cholesterol, gamma glutamyl transferase, BMI and triglycerides were associated with being classified as having dysglycaemia. Low values for chloride, creatinine and bicarbonate were associated with being classified as having dysglycaemia. Being prescribed metformin and insulin were associated with having poor glycaemic control, whereas the absence of such prescription did not tend to influence classification (figure 4).

**Figure 4:** SHAP Feature Importance Plot showing the effect of different features on model prediction



**Classified as having good glycaemic control control**

**Classified as having dysglycaemia**

**SHAP Feature Importance Plot:** Graph describing the SHAP feature importance for the CatBoost model to classify between participants based on glycaemic control. Values are based on the classification for participants in the train dataset (n = 19, 841). The Y axis shows features in order of importance from top to bottom, with random blood glucose being the most important feature. The X axis represents the mean SHAP value for that feature. Colour depicts feature value – for example, the red component of the glucose feature represents a high random blood glucose value. The midline at the zero point denotes the classification threshold, with values to the left representing participants classified as having dysglycaemia and values to the right representing paritcipants with good glycaemic control. Dysglycaemia was defined as either HbA1c ≥ 48mmol/mol or an abnormal oral glucose tolerance test (either a fasting value > 7.0 mmol/L or 2 hour result > 11.0 mmol/L).

**Clinical Implications:**

In the hold-out test set (n = 2, 331), there were 292 participants with previously recognised T2DM (12.53%), 43 participants with pre-diabetes (1.84%) and 1, 996 participants who had not previously been diagnosed with T2DM (85.63%). 115 participants in the hold-out test set had previously unrecognised diabetes (4.93% of the hold-out test set) – meaning that they had biochemical evidence suggestive of a diagnosis of T2DM but did not acknowledge a diagnosis of T2DM in the questionnaire segment of the NHANES study. The system identified 78 of the 115 participants with previously unrecognised T2DM (67.82%, figure 5). Participants with unrecognised diabetes identified by this system tended to have a more threshold diagnosis of T2DM (median HbA1c 49.72 mmol/mol), have low rates of secondary prevention and have little consideration of complications of diabetes (table 8). For example, no participant with a newly recognised T2DM was aware of their retinopathy status. This suggests that the identification of unrecognised T2DM may provide many opportunities to optimise diabetic care and improve both primary and secondary prevention for patients.

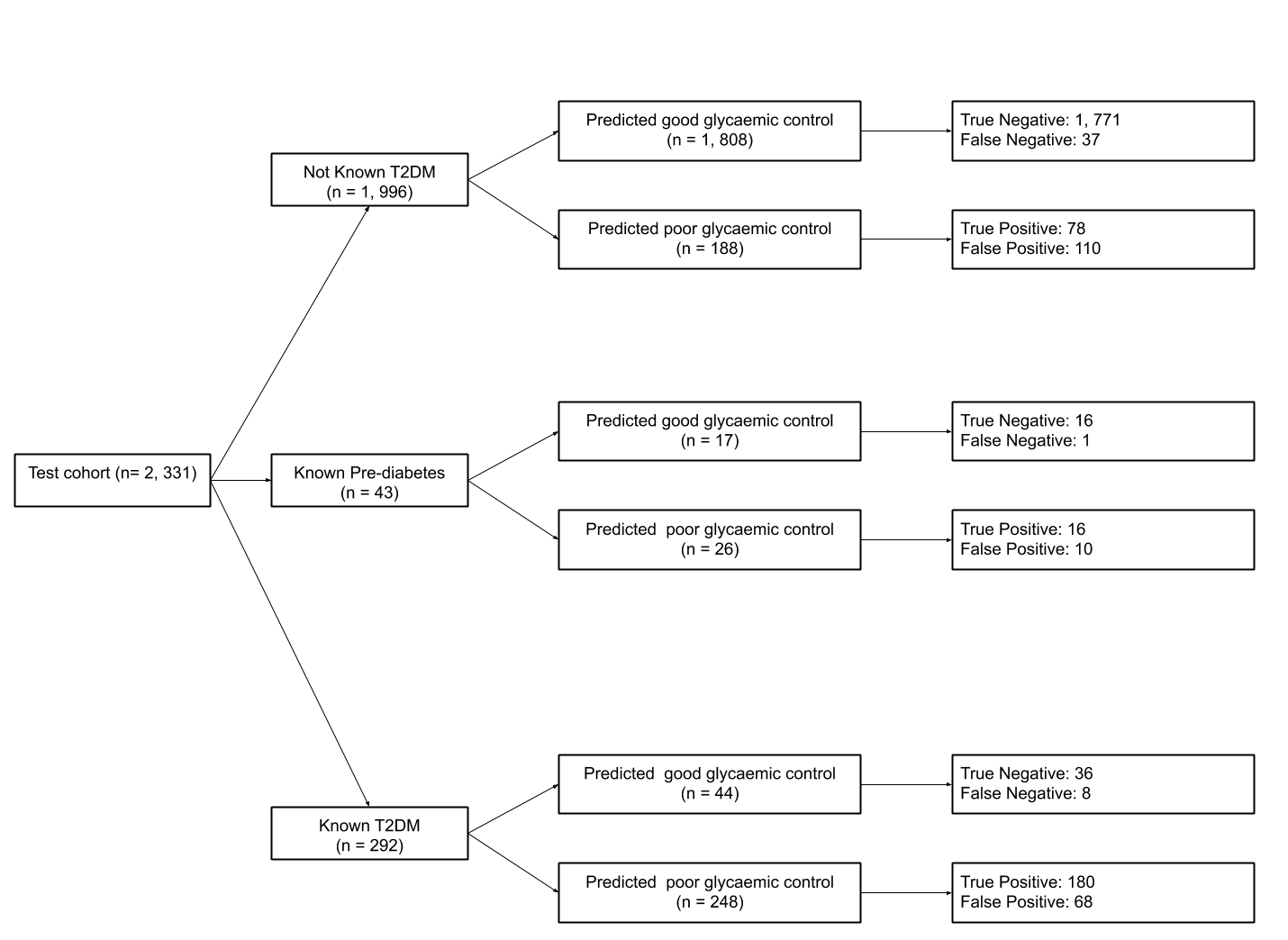
In line with the positive predictive value of the model, approximately 60% of all alerts made by the system would lead to an opportunity to improve clinical care (274 of 462 alerts, figure 5). However, it would miss 14% of all situations when an alert would be expected to be made by the system (46 of 320 occasions where there was a valid opportunity to improve clinical care).

The model had 46 participants classified as false negatives and the false negatives appear to represent a cohort with less burden of diabetic disease (table 9). The median HbA1c value for the cohort of false negative participants was significantly lower compared to the true positive cohort (43.70 mmol/mol vs. 55.18 mmol/mol, p < 0.001, table 9). Similarly, random blood glucose and triglycerides values are significantly lower in the false negative cohort and HDL cholesterol is significantly higher in the false negative cohort compared to the false negative cohort. Only 17 of these participants had an elevated HbA1c (36.96%), whereas across the whole dataset approximately 63.25% of people with biochemical evidence of dysglycaemia had an elevated HbA1c value (table 9). It appears that the model is having difficulty distinguishing and determining the classification of threshold cases.

**Table 8:** Summary table describing the participants identified by the Diabetes e-Alert System as having previously undiagnosed Type 2 Diabetes Mellitus

|  |  |
| --- | --- |
| **Feature** | **78 Participants newly identified as having T2DM when the Diabetes e-Alert System was applied to the hold-out test set (n =2, 331)** |
| **Gender**  Male  Female | 39 (50.0%)  39 (50.0%) |
| **Age** (years) | 60.0 (17.0) |
| **Ethnicity**  Non-Hispanic White       Non-Hispanic Black       Mexican American       Other Hispanic       Other race including multi-racial | 26  20  17  7  8 |
| **BMI** (kg/m2) | 32.15 (9.35) |
| **Mean arterial pressure** | 88.33 (31.9) |
| **Glycaemic measurements**  Random blood glucose (mmol/L)  HbA1c (mmol/mol) | 7.55 (2.87)  49.72 (16.4) |
| **Lipid profile**  LDL cholesterol (mmol/L)  HDL cholesterol (mmol/L) | 127.50 (58.57)  1.16 (0.41) |
| **Secondary Care**  When was the last time you saw a Diabetes Specialist? | No participant could recall or value missing |
| **Diabetic Complications**  Has the doctor ever told you that diabetes has affected your eyes or that you have retinopathy? | No participant could recall or value missing |
| **Diabetes Treatment**  Metformin  Other oral agents ±  Insulin | 4 (5.13%)  3 (3.87%)  1 (1.28%) |
| **Secondary Prevention**  Statin  ACE inhibitor/ARBs | 13 (16.67%)  11 (14.10%) |

Table 8 shows the demographics and features of the participants that were newly identified as having likely T2DM when the Diabetes e-alert system was applied to the hold-out train set. This identified 78 participants of the 115 participants with unrecognised diabetes. This cohort has threshold evidence of dysglycaemia with the median HbA1c just within a diabetic range but also a cohort that has never engaged with diabetes care previously. This is evidenced by the universal lack of knowledge about their retinopathy status. These participants have poor rates of secondary prevention. A small number of participants were already prescribed anti-diabetic therapeutics, suggesting that the self-identified diabetic status may have been wrong for some patients. Categorical variables (gender, ethnicity, medication) analysed as value counts with percentages. Continuous variables are presented as median values with interquartile ranges. ARBs: angiotensin receptor blockers. HbA1c: glycated haemoglobin. ± Other anti-diabetic agent refers to sitagliptin, liraglutide, gliclazide, gilclazide, glipizide, tolbutamide, gilbenclamide



**Figure 5:** Clinical Evaluation of Diabetes e-Alert System

Figure 5 demonstrates the clinical implications of a diabetes e-alert system. The model first “triages” patients according to whether they have a known diabetes diagnosis. This would be based on previous diagnoses coded in the electronic health record system (EHR). The model then aims to predict biochemical evidence of dysglycaemia based on routine clinical data available in a typical EHR. This is defined as a HbA1c ≥ 48 mmol/mol or an abnormal oral glucose tolerance test. The system makes 462 alerts, of which 274 alerts would have provided a clinician with a potential to improve diabetes management (sum of true positives). The model missed 46 opportunities to improve clinical care (sum of false negatives). The system would have made 188 nuisance alerts (sum of false positives) but would have detected 78 patients with previously unrecognised T2DM and 16 patients who have pre-diabetes who have poor glycaemic control.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **True Positive Cohort**  **(n = 274)** | **False Negative Cohort**  **(n = 46)** | **p value** |
| **Gender**  Male  Female | 138 (50.36%)  136 (49.64%) | 24 (52.17%)  22 (47.83%) | p = 0.8204 |
| **Age** (years) | 61.50 (19.0) | 67.0 (21.0) | p = 0.1878 |
| **Ethnicity**  Non-Hispanic White       Non-Hispanic Black       Mexican American       Other Hispanic       Other race including multi-racial | 83 (30.29%)  65 (23.72%)  58 (21.17%)  42 (15.33%)  26 (9.49%) | 23 (0.50%)  8 (17.39%)  7 (15.21%)  5 (10.87%)  3 (6.52%) | p = 0.1406 |
| **BMI** (kg/m2) | 31.60 (8.53) | 28.32 (5.76) | p = 0.1879 |
| **Glycaemic measurements**  Random blood glucose (mmol/L)  HbA1c (mmol/mol) | 7.99 (4.05)  55.18 (19.67) | 5.64 (1.0)  43.70 (11.2) | p < 0.001  p < 0.001 |
| **Diagnostic Tests**  Proportion with abnormal OGTT  Proportion with abnormal HbA1c | 99 (36.13%)  199 (72.63%) | 29 (63.04%)  17 (36.96%) | p < 0.001 |
| **Lipid profile**  LDL cholesterol (mmol/L)  HDL cholesterol (mmol/L)  Triglycerides (mg/dL) | 104.50 (61.75)  1.16 (0.44)  1.92 (1.85) | 124.0 (53.25)  1.28 (0.6)  1.59 (0.85) | p = 0.1292  p < 0.05  p < 0.05 |
| **Other Laboratory Blood Tests**  Gamma glutamyl transferase (U/L)  Chloride (mmol/L)  Bicarbonate (mmol/L)  Creatinine (µmol/L) | 26.0 (20.0)  102.0 (4.0)  25.0 (3.0)  75.14 (29.61) | 25.0 (13.25)  104.0 (4.0)  25.0 (3.75)  73.82 (34.04) | p = 0.1605  p < 0.05  p = 0.3398  p = 0.2848 |
| **Diabetes Treatment**  Metformin  Other oral agents ±  Insulin | 117 (42.70%)  54 (19.71%)  62 (22.63%) | 3 (6.52%)  2 (4.35%)  0 (0%) | p = 0.7878 |
| **Secondary Prevention**  Statin  ACE inhibitor/ARBs | 117 (42.70%)  105 (38.32% | 16 (34.78%)  14 (30.43%) | p = 0.9482 |

**Table 9:** Comparison of demographics and features between cohorts correctly and incorrectly classified by the model (true positive vs false negative).

Table 9 describes the demographics and features of participants, divided according to true positive and false negative classification in the hold-out test set (n = 2, 331). This suggests that the model had difficulty detecting threshold cases as the false negative cohort (n = 46) had significantly lower median HbA1c values, triglycerides and higher median HDL cholesterol. There was also a trend for less prescription of secondary prevention and diabetes treatment, although numbers are too small to make this a statistically significant difference. Data presented as value counts and percentages for categorical features and as median value with interquartile range for continuous variables. Data analysed using chi-square statistic for categorical variables and the Mann-Whitney test was used for continuous variables. Significance set at p < 0.05. ARBs: angiotensin receptor blockers. HbA1c: glycated haemoglobin. ± Other anti-diabetic agent refers to sitagliptin, liraglutide, gliclazide, gilclazide, glipizide, tolbutamide, gilbenclamide

**Discussion:**

Surveillance and alert systems using electronic healthcare records (EHR) have been proposed for a variety of clinical situations including acute kidney injury management12,32, early detection of clinical deterioration14, prediction of adverse events in day-case surgical patients33 and prescribing errors.15,16 The recognition and management of Type 2 Diabetes Mellitus (T2DM) could potentially be improved through the use of an e-alert system. In the present study, a prototype binary classification model was developed using a large, community-based database to predict people with poor glycaemic control (composite of either an abnormal oral glucose tolerance test (OGTT) - fasting value > 7.0 mmol/L or 2 hour result > 11.0 mmol/L, or an abnormal HbA1c ≥ 48mmol/L). Having biochemical evidence of dysglycaemia would be indicative of a new diagnosis of T2DM or poorly controlled but recognised T2DM. This diabetes e-alert system used features commonly available within a typical electronic health record system (EHR) and demonstrated good sensitivity and specificity (85.63% and 90.65% respectively), with a reasonable positive predictive value (59.31%). Alerts generated by such a system could reduce the rate of unrecognised T2DM, draw attention to poorly controlled T2DM and highlight patients who would require better primary and secondary prevention (e.g. retinopathy screening or statin therapy).

The top features influencing model classification included random blood glucose values and current diabetes treatment regimen (figure 4). Other prominent variables include known associations with metabolic syndrome (HDL cholesterol, LDL cholesterol, gamma glutamyl transferase) and known sequelae of poorly controlled T2DM such as chronic kidney disease (creatinine, bicarbonate). High values of GGT were associated with being classified as having poor glycaemic control and elevated GGT have been found to be associated with oxidative stress, non-alcoholic fatty liver disease and has been suggested as early biomarkers for the development of T2DM.34 The similarity between the features included in all final models is reassuring, suggesting that these features indeed represent the most significant features for this classification problem. Interestingly, not being prescribed a therapeutic agent to treat T2DM (either metformin or insulin) did not overly determine classification into either cohort with good or poor glycaemic control (figure 4). The model was not overly trusting of current prescribed medication to make a classification – suggesting that many participants may not be on the optimal therapy for their level of disease.

It is important to note that this model was developed in an ideal world context – a community-based, cross-sectional sample of the United States population with low rates of missing data. Further studies will be required using real-world EHR data to assess the robustness of the model to changes in study population and impurities in datasets. In the present study, approximately 10% of values were missing and these were included in the study as out-of-sample values in order to preserve some data impurity in the training process. EHR data can have a high incidence of missing-not-at-random values as the recording of some variable is determined by a clinician’s style of practice or shaped by their belief about a person’s risk of diabetes. For example, a clinician may be less likely to request an HbA1c in a person with a perceived low risk for T2DM and as a result, the model may not learn to discriminate between cases with low or high HbA1c compared to an ideal database where HbA1c values are ubiquitous. This model may also not translate to more acute medical settings where random blood glucose values can change unpredictably in the acute phase of an illness or in association with medical interventions such as steroid therapy. It is also possible that such a system may improve with real-world EHR data as missing data may in fact be informative (e.g. denoting low-risk patient) and longitudinal data (e.g. time-series laboratory tests of glucose, lipids or liver function) may provide a metric of long-term diabetic risk and wider metabolic health. Further studies will be required using real-world EHR data to assess the robustness of the model to changes in study population and impurities in datasets.

Key advantages of the present study include the large, diverse study population included in the dataset, that it is not restricted to single-centre and the careful selection of features that are realistically and reliably available in most EHRs. There has been much work done previously building classifiers to diagnose T2DM using highly-specialised features that would only be recorded in a research setting22,23. The present study utilises features that would be reliably recorded in most EHRs, meaning that missing values may actually be quite low for such features in a real-world system. Furthermore, machine learning models can be plagued by lack of representation of minority populations, which can perpetuate pre-existing health inequalities.35 Consequently, these models are biased and do not generalise to patients from black and ethnic minorities – a cohort who are particularly at risk for developing T2DM. It is likely that there was sufficient representation of minority groups in the model due to the over-sampling of traditionally under-represented groups in the NHANES methodology.

In terms of limitations, this study only assesses glycaemic control at the present time – giving an ‘instantaneous prediction’. It does not predict the likelihood of a participant going on to develop diabetes in the future, risk of developing complications of T2DM or progression to insulin therapy. However, it would be interesting to assess whether the model is predicting patients who would than progress to a diagnosis of T2DM at a later stage, particularly among the false positive cohort of participants.

It could be argued that diabetes e-alerts for poor glycaemic control may not necessarily improve patient care as it would be generating futile alerts for patients who would have been recognised as having poor glycaemic control through usual clinical practice. However, it has been shown that follow-up of abnormal results can be unreliable – with one study reporting that over 5% of people with hypothyroidism were not followed-up or prescribed thyroid hormone replacement despite clear biochemical evidence of hypothyroidism in EHR system.36 This suggests that there is a role for a system that is diligently monitoring and flagging up abnormal results as it is not possible to rely on the universal follow-up of all abnormal results.

Furthermore, the diabetes e-alert system has more merit than just considering random blood glucose values – interestingly, just 12% of participants with biochemical evidence of dysglycaemia (abnormal HbA1c or OGTT) had a high random blood glucose value. This highlights that the diabetes e-alert system can provide an aid interpreting other blood test results, drawing attention to sub-diabetic random blood sugar values in patients who may still be at high risk for T2DM. There is also a growing use of point-of-care HbA1c kits such as *A1CNow+* and *DCA* in current clinical practice. However, such tests would not negate the need for an EHR alert system as it would not be feasibly offered to all patients as a means of screening for unrecognised T2DM or poorly controlled T2DM. These point-of-care tests are also limited as clinicians must have a suspicion of T2DM in order to request such a point-of-care test. The diabetes e-alert system has the advantage over point-of-care tests and risk-factor based screening tools as the e-alert system operates silently within the EHR architecture for all patients and flags a potential opportunity to optimise T2DM care to clinicians. The diabetes e-alert system could work in tandem with point-of-care tests and trigger the use of a point-of-care HbA1c to confirm a diagnosis.

This model was developed to promote sensitivity over specificity as the principle objective of the model was to avoid a false negative result. A false negative result in this context would result in a patient with likely poorly controlled T2DM not being flagged up to the clinician. Error analysis of this model suggests that false negatives tended to be found in a cohort of patients who had milder dysglycaemia, perhaps suggesting that the model had difficulty classifying threshold cases. While this system has good levels of sensitivity, it is possible that such a high false positive rate may create an “alert fatigue” effect and result in clinicians not paying due attention to an e-alert12. Such an alert within an EHR could trigger diagnostic tests, patient education, and prescription of anti-diabetes medication or screening for complications of diabetes. Clinical trials would be required to assess how clinicians use alerts, how it influences clinical workflow and how it is perceived by clinicians and patients.

Future work could also explore whether the model could be extended to predict insulin regimen and units and act as a decision support tool within an EHR. This mimics the growing efforts to develop closed-loop systems to treat Type 1 Diabetes with insulin sensors and pumps. Insulin dose support tools for T2DM has been implemented in some tertiary care settings and have shown improvements in patient’s time within blood glucose target range and such systems were not associated with higher rates of hypoglycaemic episodes37. It would also be interesting to explore how messages and interventions from the system could be altered to target certain patient cohorts. For example, certain groups of patients may be more responsive to literature while other patients may be more amenable to structured education groups. This would mirror customer segmentation approaches common place in the business world and could act to tailor messages according to the individual patient.

**Conclusion:**

It is possible to identify patients with poor glycaemic control to a high degree of sensitivity and specificity using routinely collected clinical data. This may hold promise as a means of creating an e-alert system within an EHR to aid with the timely recognition and appropriate management of T2DM. However, further testing with sparse real-life clinical data and ultimately, clinical trials are required for a more robust assessment of such a system.

**References:**

1. 3.8 million people in England now have diabetes - GOV.UK [Internet]. [cited 2020 Apr 16]. Available from: https://www.gov.uk/government/news/38-million-people-in-england-now-have-diabetes

2. Diabetes [Internet]. [cited 2020 May 11]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes

3. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. Diabetes Res Clin Pract. 2014 Feb 1;103(2):150–60.

4. Cowie CC, Rust KF, Byrd-Holt DD, Eberhardt MS, Flegal KM, Engelgau MM, et al. Prevalence of Diabetes and Impaired Fasting Glucose in Adults in the U.S. Population National Health and Nutrition Examination. 1999;

5. Harris MI, Klein R, Welborn TA, Knuiman MW. Onset of NIDDM occurs at least 4-7 yr before clinical diagnosis. Diabetes Care. 1992 Jul 1;15(7):815–9.

6. Centre for Disease Control and Prevention. National Diabetes Statistics Report 2020 [Internet]. 2020 [cited 2020 Jun 23]. Available from: https://www.cdc.gov/diabetes/data/statistics/statistics-report.html

7. Lean ME, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. Lancet. 2018 Feb 10;391(10120):541–51.

8. Graber ML. The incidence of diagnostic error in medicine. Vol. 22, BMJ Quality and Safety. BMJ Publishing Group Ltd; 2013. p. ii21–7.

9. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak [Internet]. 2017 Feb 28 [cited 2020 May 11];17(1):24. Available from: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0418-4

10. Singh H, Giardina TD, Forjuoh SN, Reis MD, Kosmach S, Khan MM, et al. Electronic health record-based surveillance of diagnostic errors in primary care. BMJ Qual Saf. 2012 Feb 1;21(2):93–100.

11. NHS Long Term Plan [Internet]. [cited 2020 May 11]. Available from: https://www.longtermplan.nhs.uk/

12. Selby NM. Electronic alerts for acute kidney injury. Curr Opin Nephrol Hypertens [Internet]. 2013 Nov [cited 2020 May 11];22(6):637–42. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00041552-201311000-00008

13. Kucher N, Koo S, Quiroz R, Cooper JM, Paterno MD, Soukonnikov B, et al. Electronic Alerts to Prevent Venous Thromboembolism among Hospitalized Patients. N Engl J Med [Internet]. 2005 Mar 10 [cited 2020 May 11];352(10):969–77. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa041533

14. Korach ZT, Yang J, Rossetti SC, Cato KD, Kang MJ, Knaplund C, et al. Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. Int J Med Inform. 2020 Mar 1;135:104053.

15. Westbrook JI, Reckmann M, Li L, Runciman WB, Burke R, Lo C, et al. Effects of two commercial electronic prescribing systems on prescribing error rates in hospital in-patients: A before and after study. PLoS Med. 2012 Jan;9(1).

16. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Vol. 13, Nature Reviews Genetics. Nature Publishing Group; 2012. p. 395–405.

17. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Vol. 15, Computational and Structural Biotechnology Journal. Elsevier B.V.; 2017. p. 104–16.

18. Jegan C, Kumari VA, Chitra R. Classification Of Diabetes Disease Using Support Vector Machine Development of an Intelligent system for the diagnosis of cardiovascular diseases View project Classification Of Diabetes Disease Using Support Vector Machine. [cited 2020 May 11];3:1797–801. Available from: https://www.researchgate.net/publication/320395340

19. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010 Mar 22;10(1):1–7.

20. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013 Feb 1;29(2):93–9.

21. Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed. 2011 Dec 1;104(3):443–51.

22. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak [Internet]. 2019 Nov 6 [cited 2020 Jan 14];19(1):211. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31694707

23. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017 Jan 1;97:120–7.

24. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]. 2010 Mar 1 [cited 2020 May 11];2010:1–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21347133

25. World Health Organisation. Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus [Internet]. 2011 [cited 2020 Jun 23]. Available from: https://www.who.int/diabetes/publications/report-hba1c\_2011.pdf?ua=1

26. National Institute for Health and Care Excellence. Diabetes - type 2 - NICE CKS [Internet]. 2019 [cited 2020 Jun 23]. Available from: https://cks.nice.org.uk/diabetes-type-2#!scenario

27. Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987 Sep 1;27(3):221–34.

28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features [Internet]. 2018 [cited 2020 May 11]. Available from: https://github.com/catboost/catboost

29. Yao Q, Wang M, Chen Y, Dai W, Li Y-F, Tu W-W, et al. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. 2018 Oct 31 [cited 2020 Jun 23]; Available from: http://arxiv.org/abs/1810.13306

30. Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, et al. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. 2020 Mar 13 [cited 2020 Jun 23]; Available from: http://arxiv.org/abs/2003.06505

31. Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions [Internet]. 2017 [cited 2020 May 11]. Available from: https://github.com/slundberg/shap

32. Lachance P, Villeneuve PM, Wilson FP, Selby NM, Featherstone R, Rewa O, et al. Impact of e-alert for detection of acute kidney injury on processes of care and outcomes: Protocol for a systematic review and meta-analysis. Vol. 6, BMJ Open. BMJ Publishing Group; 2016. p. e011152.

33. Kaafarani HMA, Rosen AK, Nebeker JR, Shimada S, Mull HJ, Rivard PE, et al. Development of trigger tools for surveillance of adverse events in ambulatory surgery. Qual Saf Health Care [Internet]. 2010 Oct 1 [cited 2020 Jun 23];19(5):425–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20513790

34. Lee DH, Ha MH, Kim JH, Christiani DC, Gross MD, Steffes M, et al. Gamma-glutamyltransferase and diabetes - A 4 year follow-up study. Diabetologia [Internet]. 2003 Mar 1 [cited 2020 Jun 24];46(3):359–64. Available from: https://link.springer.com/article/10.1007/s00125-003-1036-5

35. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. Vol. 178, JAMA Internal Medicine. American Medical Association; 2018. p. 1544–7.

36. Schiff GD, Kim S, Krosnjar N, Wisniewski MF, Bult J, Fogelfeld L, et al. Missed Hypothyroidism Diagnosis Uncovered by Linking Laboratory and Pharmacy Data. Arch Intern Med [Internet]. 2005 Mar 14 [cited 2020 Jun 23];165(5):574. Available from: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.165.5.574

37. Neubauer KM, Mader JK, Höll B, Aberer F, Donsa K, Augustin T, et al. Standardized Glycemic Management with a Computerized Workflow and Decision Support System for Hospitalized Patients with Type 2 Diabetes on Different Wards. In: Diabetes Technology and Therapeutics. Mary Ann Liebert Inc.; 2015. p. 685–92.