LUND UNIVERSITY
Faculty of Medicine

Master's Programme in Public Health

Predictive modelling using a nationally representative database to identify the determinants of prediabetes; a machine learning analytic approach on the National Health and Nutrition Examination Survey (NHANES) 2013-2014

Spring 2018

Author
Kushan Kumara De Silva Ranakombu

Supervisors
Daniel Jönsson, Associate professor, Faculty of Odontology, Malmo University, Sweden
&
Ryan Demmer, Associate professor, Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, USA

ABSTRACT

Background: Prediabetes is a global epidemic with rising prevalence rates, but its diagnosis based on traditional risk factors is challenging. Application of novel machine-intelligence based methods to public health databases could provide valuable insights into the disease process.

Aim: To build predictive models to elucidate the determinants of prediabetes using machine learning algorithms on a nationally representative sample of the US population.

Method: Two datasets containing general (n = 6346) and dental (n = 3167) variables were prepared from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 and were randomly partitioned to create train and internal validation data. Feature selection algorithms were run on the train (n = 3174) data containing 156 pre-selected general variables. Five machine learning algorithms were applied on train data containing general (n = 3174) and dental (n = 1584) variables as well as on re-sampled datasets built using 4 resampling methods. Predictive models were tested on internal validation data containing general (n = 3172) and dental (n = 1583) variables. External validation was done on 2 datasets containing general (n = 3000) and dental (n = 1500) variables prepared from the NHANES 2011-2012. Model performance was evaluated using area under the receiver operating characteristic curve (AUC). Determinants were elucidated by odds ratios in logistic regression models and by variable importance values in other algorithms. The CDC prediabetes screening tool was chosen as the benchmark against which the performance of optimal models was compared.

Results: Seven optimal (>70% AUC) models built on the dataset containing general variables elucidated 25 determinants of prediabetes including a few novel associations; 20 were identified by both logistic regression and other non-linear/ensemble models while 5 were solely elucidated by the latter. Dental variables by themselves were not predictive of, and periodontitis appeared the only dental determinant of, prediabetes. The optimal machine learning model (AUC = 71.6%) built on the data containing general variables outperformed the chosen benchmark while that built on dental data equaled the performance of the screening tool.

Conclusion: A range of determinants of prediabetes was identified through validated and benchmarked models highlighting the potential of a systematic, machine intelligence-based modelling approach on a public health database to elucidate the determinants of prediabetes including novel predictors.

Keywords: prediabetes, determinants, machine learning, feature selection, NHANES

DEFINITIONS

Machine learning	A discipline focused on developing computer algorithms for transforming data into intelligent action by the concomitant use of both computing power and statistical methods (Lantz, 2013).
Predictive analytics (predictive modelling)	Application of statistical methods including machine learning algorithms on both new and historical data to make predictions about future or unknown events (Strickland, 2015)
Data mining	Application of machine learning, statistics, and informatics methods on large amounts of data to extract new knowledge (Zhao, 2012)
Supervised learning	Learning process involving the modelling of a set of input (independent) variables and an output (dependent) variable and the prediction on new data using constructed models (Tripathi, 2017)
Unsupervised learning	Learning process involving the modelling of input data without response variables to explore patterns and statistical structures in the data (Tripathi, 2017)
Classification and regression tree (CART/Decision tree)	A machine learning algorithm comprised of a set of split conditions aiming for accurate prediction (continuous response variable) or classification (categorical response variable) of cases (Loh, 2011)
Bagging (Bootstrap aggregating)	A machine learning ensemble meta-algorithm to improve accuracy and stability and reduce variance. Multiple bootstrap (i.e. sampling with replacement) train samples are used to train different models and the model outputs are combined by averaging (in regression) or voting (in classification) to create a single output (Breiman, 1996).
Boosting	A machine learning ensemble meta-algorithm to improve accuracy, reduce bias (and also, variance

	in supervised learning) that convert weak learners to a strong prediction rule (Schapire, 2003)
Bagged CART	An ensemble learner in which bagging/bootstrap aggregating is applied to decision tree algorithm. Multiple bootstrap sub-samples of train data are created, a decision tree model is trained on each sample, and a single output is produced by averaging (in regression tree) or voting (in classification tree) (Breiman, 1996)
Gradient boosting	An ensemble machine learning algorithm used for classification and regression that consists of boosting and optimization of a loss function (Friedman, 2001; Friedman, 2002)
Random forest	An ensemble machine learning algorithm used for classification and regression in which a large number of decision trees are constructed during the training process and the mode of the predicted classes (in classification) or the mean prediction of individual trees (in regression) is produced as output (Breiman, 2001)
Artificial neural network	A non-linear machine learning algorithm modelled on the human brain and nervous system that learns from and adapts to initial inputs. An artificial neural network model generally consists of a set of adaptive weights (numeric parameters tuned by a learning algorithm) and can be approximated by a non-linear function of its inputs (Strickland, 2015)
Ensemble learning	A machine learning method in which many classifiers are generated, and their outputs are aggregated. Most frequently used methods are based on bagging and boosting (Zhang & Ma, 2012).
Resampling	Restructuring a dataset having class imbalance through an iterative random sampling process to

	obtain more balanced data for training machine learning algorithms. Has been proven to be an effective strategy for training imbalanced data (Jagelid & Movin, 2017)
Random oversampling (ROSE)	A resampling method by which the prevalence of the minority class is increased in the data (Lauron & Pabico, 2016)
Synthetic minority oversampling technique (SMOTE)	A resampling method in which a combination of over-sampling the minority class and under-sampling the majority class is employed to achieve optimal classifier performance (Chawla et al, 2002)
Class imbalance	The situation in which a dataset overrepresents one class and under-represents the other class. This poses a problem in machine learning because many algorithms tend to bias the majority class and ignore/misclassify the minority class (Longadge & Dongre, 2013)
Confusion matrix	A table illustrating the performance of a machine learning algorithm. Rows of the matrix represent the instances in predicted classes and columns represent the instances in actual classes (or vice versa) (Lesmeister, 2017)
Receiver operating characteristic (ROC) curve	A graph illustrating the predictive ability of a binary classifier at varying discrimination thresholds. Obtained by plotting true positive rate (sensitivity) of a classifier against its false positive rate (1-specificity) (Lesmeister, 2017)
Area under the ROC curve (AUC)	A robust indicator of a machine learning algorithm's predictive performance preferred in imbalanced data training (Chawla, 2009)
Accuracy	A confusion matrix metric indicating how correct overall a given classifier's prediction performance is.

	Accuracy = (number of true positives + number of true negatives)/number of total predictions
Sensitivity	True positive rate; the proportion of positives that are correctly identified as such Sensitivity = number of true positives/number of total positives = number of true positives/ (number of true positives + number of false negatives)
Specificity	True negative rate; the proportion of negatives that are correctly identified as such Specificity = number of true negatives/number of total negatives = number of true negatives/ (number of true negatives + number of false positives)
Negative predictive value (NPV)	The proportion of negative results that are true negative results. Indicates the probability that subjects with a negative prediction outcome are truly outcome-negative. NPV = number of true negatives/ (number of true negatives + number of false negatives)
Positive predictive value (PPV)	The proportion of positive results that are true positive results. Indicates the probability that subjects with a positive prediction outcome truly have the outcome. PPV = number of true positives/ (number of true positives + number of false positives)
Kappa	A measure of how well the classifier performed as compared to how well it would have performed simply by chance. Kappa = (Observed agreement - chance agreement)/ (1-chance agreement)
Multivariate imputation by chained equations (MICE) algorithm	A proven method of multiply imputing missing data which can handle both continuous and categorical variables as well as complex missing data structures (Azur et al, 2011)

Feature selection	The strategies applied for selecting a subset of relevant attributes/variables/predictors for building statistical or machine learning models. Objectives are to make simple, parsimonious models, reduce training time and enhance predictive ability (Guyon & Elisseeff, 2003)
Variable importance	An evaluation metric of machine learning models indicating the relative contribution/usefulness of each variable for predicting the response variable and are analogous to the concept of statistical significance. Methods of calculation may differ by algorithms and packages used but there is no one "best" measure neither a single definition (Grömping, 2009)

ABBREVIATIONS AND ACRONYMS

AA degree……	Associate of Arts degree
ADA……	American Diabetes Association
ALT……	Alanine amino transferase
AMT……	Aspartate amino transferase
ANN…….	Artificial neural network
AUC……...	Area under the receiver operating characteristic curve
Bagged CART……...	Bootstrap aggregated classification and regression tree
BMI……	Body mass index
CDC……...	Centers for Disease Control and Prevention
CI……...	Confidence interval
CVD……...	Cardiovascular disease
DBP……	Diastolic blood pressure
DM ……	Diabetes mellitus
DPP4……..	Dipeptidyl peptidase-4
ERB……	Research Ethics Review Board
FPG……	Fasting plasma glucose
GB……..	Gradient boosting
GGT……...	Gamma glutamyl transferase
GLM……..	Generalized linear model
HbA1c……	Glycated hemoglobin
HBV……...	Hepatitis B virus
HCV……...	Hepatitis C virus
HIIT……...	High intensity interval training
IgG…….	Immunoglobulin G
MICE…….	Multivariate imputation by chained equations algorithm
MICT…….	Moderate intensity continuous training
NCHS…….	National Center for Health Statistics
NHANES……...	National Health and Nutrition Examination Survey
OGTT…….	Oral glucose tolerance test
OR……..	Odds ratio
RBC……	Red blood cell count
RF……...	Random forest
ROC……...	Receiver operating characteristic analysis/curve
ROSE…….	Random oversampling algorithm
SBP……	Systolic blood pressure
SD……..	Standard deviation
SMOTE……..	Synthetic minority oversampling technique
US……...	United States
WBC……...	White blood cell count
WHO……...	World Health Organization

CONTENT

1. INTRODUCTION

1.1. Background

Prediabetes is a global epidemic, the prevalence of which is on the rise across the world (Edwards et al., 2016). It encompasses the high-risk pool in the intermediate, reversible stage between normal glucose tolerance and overt diabetes mellitus (DM) which deserves unequivocal attention as it presents a window of opportunity to curtailing progression to established diabetes and reverting to normal blood glucose levels by receiving timely, evidence-based care. Prediabetes management strategies frequently consist of simple lifestyle interventions, and less frequently, pharmacotherapy and bariatric surgery (Bansal, 2015).

Meta-analyses have provided high levels of evidence for the complications associated with prediabetes including cardiovascular disease (CVD) (Huang et al., 2016), stroke (Lee, M. et al., 2012), chronic kidney disease (Echouffo-Tcheugui et al., 2016), cancer especially liver, endometrial, stomach/colorectal (Huang et al., 2014), and pancreatic (Fu et al., 2016). Socio-demographic parameters such as poverty and older age (Mainous et al., 2014), clinical parameters such as obesity, hypertension, periodontitis (Demmer et al., 2015; Andriankaja et al., 2014) hypertriglyceridemia (Hilawe et al., 2016), arthritis (Okwechime et al., 2015) and biochemical parameters such as dipeptidyl peptidase-4 (DPP4) (Zheng et al., 2014), mean serum uric acid (Zhang et al., 2016), adiponectin (Jiang et al., 2016), c-reactive protein (Jaiswal et al., 2012) and serum ferritin (Sharifi et al., 2008) have been associated with prediabetes as concluded by studies done in various geographic settings among different study groups.

1.2. The database

The national health and nutrition examination survey (NHANES) is a continuous epidemiological study designed to assess the health and nutrition of the population of the United States (US). The survey is unique in that it combines interviews and physical examinations and provides a comprehensive database for epidemiological research comprised of a nationally-representative sample of the US population. The NHANES provides interview data of demographic, socioeconomic, dietary, and health-related parameters, examination data on medical, dental, and physiological measurements, as well as data on various laboratory tests

administered by trained professionals. Therefore, it provides a unique opportunity to perform high-powered predictive analytics on a nationally-representative database to determine the predictors of various diseases. For example, Villa-Zapata et al. (2016) performed risk modelling to determine the predictors of microalbuminuria using regression-based techniques on the NHANES 2007-2008.

The NHANES database provides data to build prediction models using dental variables as well. In fact, several studies revealed that dental parameters by themselves could also be used to identify prediabetes in a dental clinic setting (Lalla et al., 2011; Holm et al., 2016) although their predictive potential in a public health database has not been evaluated to date. On the other hand, dental variables such as periodontal disease which is known as the sixth complication of diabetes (Löe,1993) and demonstrates a strong, bidirectional relationship with diabetes (Grossi and Genco, 1998; Lalla & Papapanou, 2011), and the number of missing teeth (Lalla et al., 2011) could be potential indicators of precursor stages of diabetes and their validity in a community setting for diagnosing prediabetic individuals has neither been adequately explored.

1.3. Diagnostic tests of prediabetes

The NHANES contains data on 3 diagnostic tests for prediabetes/diabetes; fasting plasma glucose (FPG), oral glucose tolerance test (OGTT) and glycated hemoglobin (HbA1c). A concomitant use of multiple prediabetes/diabetes tests has been found more effective than using a single test (Aekplakorn et al., 2015; Ghazanfari et al., 2010). The HbA1c test, universally considered the gold standard of chronic glycemic control, is not only a standard diagnostic marker of prediabetes and diabetes but also a robust prognostic indicator of diabetic complications. For instance, HbA1c levels of prediabetic range are associated with a higher risk of diabetes and CVD even after adjusting for fasting glucose and key CVD risk factors (Selvin et al., 2010). Along the continuum of the natural history of the disease, HbA1c therefore has important implications as a diagnostic tool of prediabetes (Warren et al., 2017). Nevertheless, its performance seems highly context-specific which performed poorly (Hellgren et al., 2017; Nowicka et al., 2011) and was inferior to both OGTT and FPG tests (Dong et al., 2011; Ghazanfari et al., 2010) in several settings. On the other hand, using more than one diagnostic criterion yielded superior performance ((Kim et al., 2015).

Differences prevailing in the diagnostic criteria for prediabetes recommended by the World Health Organization (WHO) and the American Diabetes Association (ADA) (Bansal, 2015) are worthy of note. The WHO cut-off figures for diagnosing prediabetes are 110-125 mg/dl for FPG and 140-200 mg/dl for OGTT (World Health Organization, 2006). Corresponding ADA cut-offs for OGTT is the same but the range for FPG is 100-125 mg/dl. Moreover, the ADA diagnostic criteria have included a range of 5.7%-6.4% HbA1c as indicative of prediabetes (American Diabetes Association, 2014).

1.4. Public health relevance

From a public health perspective, elucidating the predictors of impaired glucose control (i.e., prediabetes) at an early stage prior to diabetes development can pave the way for timely identification of the high-risk pool across the disease spectrum. This in turn may facilitate the design of early and effective public health interventions aimed at high-risk individuals before the onset of diabetes addressing the factors associated with their increased susceptibility to demonstrate persistently poor glycemic control. Prediabetes, though essentially precedes the onset of overt diabetes, frequently exists an asymptomatic condition, posing a diagnostic challenge (Bansal, 2015).

The primary aim of prediabetes screening tools is to diagnose high-risk individuals before the onset of diabetes and both ADA and CDC screening tools were found to perform robustly among the US population (Poltavskiy et al., 2016). Therefore, these screening tools may act as feasible benchmarks to be compared against the performance of any newly-constructed predictive models. For instance, a standard screening tool validated for the Korean population (Lee, Y.H. et al., 2012) was used for benchmarking machine learning models for prediabetes built from a Korean public health database (Choi et al., 2014).

Early diagnosis of prediabetes through multifactorial risk scores is a cost-containing strategy highly beneficial to any health system since relatively low-cost lifestyle modifications are the cornerstone of its management that may reduce the relative risk of developing diabetes by 40%-70% (Tabá et al., 2012). Current prediabetes risk assessment tools based on a broadly similar set of risk factors could be enhanced by incorporating novel attributes such as hematological and metabolic traits (Rathmann et al., 2010; Wilson et al., 2007). An entire health system may be

benefited by an effective diagnostic strategy of prediabetic individuals as it will cut down the exorbitant costs associated with overt diabetes management.

1.5. Analytic approach: machine learning

Both descriptive (association and clustering) and predictive (classification and regression) data-mining methods, also known as unsupervised and supervised machine learning techniques respectively, are increasingly being applied in big data based public health research using large datasets. Though more complex than traditional statistical analyses, such data mining approaches are often preferred due to their merits such as the ability to generate new hypotheses and provide novel insights into various diseases (Kavakiotis et al., 2017). A systematic review of data mining technologies for diabetes revealed that they were applied in domains such as the prediction and diagnosis of diabetes and diabetic complications, feature selection, health care flow analysis, adverse drug effect analysis, clinical guidelines enrichment, prediction of early mortality, epigenetic and genomic data analysis and biomarker identification (Marinov et al., 2011). A summary review of studies of prediction models applied in prediabetes and associated conditions is presented in the **appendix: table 1**.

1.6. The need for additional strategies

A systematic review of the quality of predictive models on diabetes revealed that the most common methodological drawbacks were univariate pre-screening of variables, categorization of continuous attributes and poor handling of missing data (Collins et al., 2011). Another systematic review of prediabetes risk assessment tools identified additional methodological issues such as the lack of external validation and calibration of tools (Barber et al., 2014). Also, it has been reported that single feature selection methods are likely to be more biased than ensembles (Neumann et al., 2017). Class imbalance is a common phenomenon in medical databases which can heavily deteriorate classifier performance as they tend to optimize the overall accuracy without considering the relative distribution of each class (Mazurowski et al., 2008; Li et al., 2010; Lusa, 2010), and several techniques such as resampling methods (Li et al., 2010; Rahman & Davis, 2013) and ensemble learning (Han et al., 2015; Alghamdi et al., 2017) have been suggested to circumvent this problem.

Since the reliability and accuracy of a prediction model is dependent on the methodological quality, rigorous steps would be needed to overcome these common methodological flaws which may include the use of advanced feature selection tools for dimension reduction, plausible construction of composite variables, sound missing data management techniques, external validation and benchmarking of models as well as the use of strategies such as resampling methods and ensemble learners to overcome the class imbalance issue.

In this context, the present study addressed the following aim and objectives:

1.7. Aim and objectives

The study aimed at building and validating prediction models to elucidate general and dental determinants of prediabetes using machine learning algorithms on a public health database. Specific objectives were:

1. To determine the general (socio-economic, clinical, and biochemical) predictors of prediabetes by a machine learning based predictive modelling approach using the NHANES 2013-2014 database
2. To determine the dental predictors of prediabetes by a machine learning based predictive modelling approach using the NHANES 2013-2014 database
3. To evaluate the predictive modelling performance of machine learning models that were constructed using general and dental parameters against a national benchmark for prediabetes screening; the CDC prediabetes screening tool

1.8. Research questions

Aligned to the above, the present study addressed the following research questions

1. Can a machine learning based predictive modelling approach on the NHANES effectively identify general (socio-economic, clinical, and biochemical) and dental determinants of prediabetes?
2. How would a machine learning based predictive modelling approach on the NHANES perform, compared to a standard prediabetes screening tool applied among the US population?

2. METHOD

Analyses were done using R statistical software and its assortment of machine learning packages. A brief description of the main R statistical packages used is given in the **appendix: table 2**. Study population comprised of participants of the NHANES 2013-2014; the latest data available in the NHANES database. The methodological approach is illustrated in the **appendix: figure 1**.

2.1. Construction of the binary outcome variable i.e. prediabetic versus normoglycemic

The binary outcome variable i.e. prediabetic versus normoglycemic in the present study was defined integrating all 3 diagnostic tests available in the NHANES database, namely, FPG, OGTT and HbA1c tests, which is commensurate with the approach adopted by Ogunyemi et al. (2015) and Zhang et al. (2015). The NHANES 2013-2014 database provided HbA1c measurements for a full sample as well as FPG and OGTT results for sub-samples. Standard reference ranges for prediabetic, diabetic and normoglycemic statuses recommended by the ADA are as follows: HbA1c: normoglycemic <5.7%, pre-diabetic = 5.7%-6.4%, diabetic>6.4%; FPG: normoglycemic <100 mg/dl, pre-diabetic = 100-125 mg/dl, diabetic>125 mg/dl; OGTT: normoglycemic <140 mg/dl, prediabetic = 140-200 mg/dl, diabetic >200mg/dl (American Diabetes Association, 2014).

Since the study was based on a nationally-representative sample of the US population, standard prediabetes diagnostic criteria recommended by the ADA were used. Individuals with HbA1c >6.4% or FPG >125 mg/dl or OGTT >200mg/dl indicating prevalent diabetes were first excluded. Of the remaining sample, participants were classified as prediabetic if they met at least one of the following criteria; FPG 100-125 mg/dl, OGTT = 140-200mg/ dl, HbA1c = 5.7-6.4%. Thus, a dichotomous outcome variable was created, premised on the agreed cut-off levels of each; prediabetic (HbA1c = 5.7% - 6.4% or FPG = 100-125mg/dl or OGTT = 140-200 mg/dl) versus normoglycemic otherwise.

Self-reported prediabetes data were also available in the NHANES database in the form of the responses to the question “Have you ever been told by a doctor or other health professional that you have prediabetes?” A receiver operating characteristic (ROC) analysis of self-reported status of having ever been told prediabetic versus current HbA1c, FPG, and OGTT among the study sample was carried out to assess the feasibility of using these data for the definition of the

outcome variable. Based on the ROC analysis, self-reported prediabetes data were excluded from outcome definition (figure 1).

2.2. Preparation of data

Two different datasets from the NHANES 2013-2014 were constructed for modelling with general and dental variables as described below. Also, two corresponding datasets with relevant general and dental variables were constructed from the NHANES 2011-2012 for the external validation of optimal models.

2.3. Variable handling, missing data management, and the construction of train and validation datasets for objective 1 (modelling with general socio-economic, clinical, physiological, and biochemical variables)

Variables with 30% or more missing data were excluded. From the repertoire of variables in the NHANES 2013-2014, 156 variables were pre-selected following a comprehensive literature review on the determinants of prediabetes. The list of pre-selected variables and the evidence/rationale for their inclusion is given in the **appendix: table 3**. Of note, the list contained a few newly created/modified/ composite variables and information about their construction is also provided in the **appendix: table 3**. A "missing at random" pattern of the missing data distribution was assumed and hence a multiple imputation was deemed feasible. Default functions of the "mice" (multivariate imputation by chained equations) package in R (Buuren & Groothuis-Oudshoorn, 2010) were used for multiply imputing missing values of each type of variable; predictive mean matching for numeric, Bayesian polynomial logistic regression for multi-level (>2 levels) categorical, and Bayesian binary logistic regression for dichotomous categorical variables, respectively. Goodness of fit of the imputed data was evaluated by comparing summary measures and distributions of variables in the original and complete datasets (**appendix: tables 4 & 6**).

A random 50/50 partitioning of the processed, complete dataset from NHANES 2013-2014 was done to create a train dataset for modelling (N = 3174) and an internal validation dataset (N = 3172). In addition, a random sample with corresponding variables was created from the NHANES 2011/2012 database for external validation of the constructed models (N = 3000). Missing data of the external validation dataset were handled similarly to the method used for the

NHANES 2013-2014 described above using the multiple imputation functions of the "mice" package. Of note, one variable, namely "diagnosed jaundice", was not available in the NHANES 2011-2012 and a random, simulated sample of values from the NHANES 2013-2014 data of this variable was added to the external validation dataset.

2.4. Feature selection for objective 1 (modelling with general socio-economic, clinical, physiological, and biochemical variables)

In accordance with the fundamental principle of feature selection that it should be performed on the data that are not used for internal or external validation of models (Guyon & Elisseeff, 2003), feature selection algorithms were run only on the train data. Thus, an array of feature selection algorithms encompassing all 3 types - wrapper, filter and embedded - was run on the train dataset (N = 3174) containing the 156 pre-selected variables. A summary of feature engineering algorithms employed, and the features extracted by each method are given in **table 2**. Based on the results of feature engineering as well as the comprehensive literature review, 46 variables that would minimize redundancy were selected for modelling. When several similar or comparable variables had appeared in the output, objectively-measured physiological or biochemical variables were selected in favor of those emanating from self-reported data. A descriptive summary of the selected 46 features is given in **table 1.1.**

2.5. Modelling and model validation process for the objective 1

Five algorithms were used to encompass linear, non-linear and ensemble models as follows:

1. Logistic regression (linear)
2. Artificial neural network (ANN) (non-linear)
3. Random forests (RF) (ensemble)
4. Gradient boosting (GB) (ensemble)
5. Bootstrap aggregated classification and regression tree (Bagged CART) (ensemble)

To address the issue of class imbalance inherent in the data, 2 methods endorsed in literature were used; 3 ensemble models as mentioned above ((Han et al., 2015; Alghamdi et al., 2017) and resampling techniques (Li et al., 2010; Rahman & Davis, 2013). Thus, in addition to the 3 ensemble algorithms mentioned above, 4 resampling techniques, namely, under-sampling, oversampling, random oversampling (ROSE) (Lunardon et al., 2014) and synthetic minority

oversampling (SMOTE) (Chawla et al., 2002) were employed. Therefore, using each of the 5 algorithms, 5 models were built as follows: a model with;

1. Original data
2. Resampled data using under-sampling
3. Resampled data using oversampling
4. Resampled data using ROSE
5. Resampled data using SMOTE.

Parameter tuning and five-fold cross-validation were performed for the ANN models while the other 4 algorithms were trained using default parameters and ten-fold cross-validation and the methodological details are given in the **appendix: table 2**. The resulting 25 machine learning models built on the train data consisting of extracted general predictors were tested on both the internal and external validation data. Predictive accuracy of models on validation data was gauged via confusion matrix metrics and area under the ROC curve (AUC). The relative impact of predictors in logistic regression models was gauged via adjusted odds ratios (OR), their confidence intervals (CI), and corresponding p-values while the variable importance values were used for elucidating predictors in case of the other 4 classification algorithms. Default, in-built functions available in the R statistical packages were used to calculate the variable importance estimates and methodological details are given in the **appendix: table 2**. Owing to the class imbalance of the sample (prevalence of prediabetes of the sample = 23.43%), AUC was chosen as the model performance evaluation criterion (Chawla, 2009; Kotsiantis et al., 2006). The cut-off AUC and other performance evaluation metrics that maximized the Youden index (sensitivity + specificity – 1) were determined for each model. The Youden index was chosen as it is preferred in imbalanced datasets (Bekkar et al., 2013). Thus, a benchmark AUC of 70% which had been endorsed as an acceptable prediction level (Jayanthi et al., 2017) was set, and 7 optimal models exceeding the benchmark were identified. A summary of these optimal models and determinants elucidated are given in **table 3.1.** and **table 3.2**.

2.6. Variable handling, missing data management, and the construction of train and validation datasets for objective 2 (modelling with dental variables)

Similar to the method adopted for the objective 1, variables with 30% or more missing data were excluded. Dental variables available in the NHANES 2013-2014 were selected via a

comprehensive literature survey on their association with prediabetes. Owing to the fewer number of variables (i.e. 15), feature selection was not performed. Details about the selected/constructed composite variables, their definition/construction and the rationale for their inclusion are given in the **appendix: table 5** while a descriptive summary is given in **the table 1.2**. Of note, the study sample comprised only of adult >= 30 years of age. This was deemed necessary because a number of dental variables, namely, periodontitis, dental floss use frequency, mouthwash use frequency, ever having a periodontal treatment, and self-reported tooth mobility, had been measured only on individuals of 30 years of age or older. This approach was further supported by the skewed distribution of prediabetes; nearly 81% of the prediabetic individuals in the sample consisting of general predictors (N= 6346) were 30 years or older. Missing data were multiply imputed via the same method adopted for the objective 1 and post-hoc accuracy analyses of the imputation was performed (**appendix: table 6**). A random 50/50 partitioning of the processed, complete dataset from NHANES 2013-2014 was done to create a train dataset for modelling (N = 1584) and an internal validation dataset (N = 1583). In addition, a random sample was created from the NHANES 2011/2012 database for external validation of the constructed models (N = 1500). Missing data of the external validation dataset were also multiply imputed using the "mice" algorithm described above.

2.7. Model-building and validation process for the objective 2

The process was similar to that which was adopted for the objective 1. Thus, 5 models, 1 with original data and 4 with re-sampled data, were built using each of the 5 algorithms resulting in 25 models in total. Parameter tuning and five-fold cross-validation were performed for the ANN models while the other 4 algorithms were trained using default parameters and ten-fold cross-validation and the methodological details are given in the **appendix: table 2**. Owing to the class imbalance of the sample (prevalence of prediabetes of the sample = 29.46%), AUC was chosen as the model performance evaluation criterion. The cut-off AUC and other performance evaluation metrics that maximized the Youden index (sensitivity + specificity – 1) were determined for each model. Since none of the 25 models exceeded the chosen benchmark AUC of 70% which had been endorsed as an acceptable prediction level (Jayanthi et al, 2017), a summary of the optimal models having the highest AUC built using each algorithm and determinants elucidated are given in **table 4**. The relative impact of predictors was gauged via

adjusted OR, CI and corresponding p-values in the case of logistic regression models and via variable importance values in case of other classification algorithms. Default, in-built functions available in the R statistical packages were used to calculate variable importance estimates and details are given in the **appendix: table 2**.

2.8. Benchmarking of predictive models (objective 3)

Predictive performance estimates of constructed optimal models under objectives 1 & 2 which had the highest AUC on the internal and external validation data were compared against the performance of a national benchmark using standard test metrics (i.e. accuracy, sensitivity, specificity, kappa, and AUC). A screening tool endorsed for the US population by the Centers for Disease Control and Prevention (CDC), namely, the CDC prediabetes screening test (CDC prediabetes screening test, 2018) was adapted for benchmarking as the study data were from a nationally-representative sample of the US population. The CDC prediabetes screening tool consists of 7 questions pertaining to age, having delivered an overweight baby (>9lb), siblings or parents having diabetes, physical activity, and obesity. The total score ranges 0-18 and the cut-point for prediabetes is 9. Adaptation of the CDC screening tool to be able to use on the NHANES data and the allocation of corresponding scores was as per Poltavskiy et al. (2016) and details are presented in the **table 6**. Age was categorized with the cut-points of 45 and 65. Since the NHANES 2013-2014 did not collect family history of diabetes information separately for parents and siblings, the 2 questions on the parents' and siblings' diabetes were combined and assigned the score of 2. Classification of obesity as per the given weight and height chart corresponds to BMI cut-point of 27 kg/m^2 and thus two categories were created and the obese were allocated a score of 5. For "physical activity", a binary variable was created based on if any of the following activities were done 5 or more days in a typical week: vigorous or moderate work, recreational work, walk or bicycle and were given different scores considering both the age and the physical activity, as shown in the **table 6**. A score of 1 was given if a female reported of having an overweight baby at birth (>9lb). Individuals with a total score ≥ 9 were categorized as pre-diabetic and those with <9 non-prediabetic. This classification was performed on both internal and external validation datasets containing general and dental variables, the performance evaluation metrics were calculated and were compared with those of optimal models built under the objectives 1 & 2. For statistical comparisons of predictive performance of optimal models

against the benchmark, the test for comparing AUC of two ROC curves by Hanley and McNeil (1982) was used. Findings are summarized in **table 5**.

2.9. Ethical considerations

Ethical approval for the NHANES 2013-2014, specified as "Continuation of Protocol #2011-17", was obtained from the National Center for Health Statistics (NCHS) Research Ethics Review Board (ERB) (National Center for Health Statistics, 2018). Strictest confidence of the collected data was guaranteed, and the written informed consent process consisted of assuring that the collected information would be used only for stated purposes and would not be rendered accessible to third parties without the prior consent of the survey participants. Publicly-available NHANES 2013-2014 data have been anonymized. Rigorous measures to ensure safety and privacy of participants were taken since the NHANES involved collection, storage, and retrieval of biological specimens and the use of invasive data collection procedures. Iatrogenic risks were minimized by deploying trained personnel for data collection. Since the present study entailed a secondary analysis of open-access, anonymized survey data, ethical approval was not required.

3. RESULTS

3.1.Analysis of predictive value of self-reported prediabetes

The Youden index-maximizing AUC estimates of ROC curves illustrating the diagnostic ability of self-reported status of ever being told prediabetic versus current Hb1Ac (**figure 1: plot a**), FPG (**figure 1: plot b**) and OGTT (**figure 1: plot c**) levels were 66.1%, 64.7%, and 65.7% respectively, and since they were thus below the 70% benchmark AUC indicating sub-optimal predictive performance, self-reported prediabetes data were not used for defining prediabetes.

3.2. Post-hoc analyses of the multiple imputation algorithm's robustness

Distributional analyses of the imputed and original datasets containing general variables revealed that out of the 156 attributes, only 1 numeric variable, namely, processed food expenditure (missing data percentage =1.61%), and 6 categorical variables, namely, having ever served in the armed forces (missing data percentage = 14.36%), marital status (missing data percentage = 21.87%), self-reported kidney stones (missing data percentage = 21.92%), past any tobacco use, (missing data percentage = 22.90%), self-reported urinary leakage (missing data percentage = 28.88%), and functional limitations (missing data percentage = 21.82%) demonstrated significantly different distributions between the 2 datasets(**appendix: table 4**). There were no significant differences in the distributions of any of the numerical and categorical variables between original and imputed datasets containing dental attributes (**appendix: table 6**).

3.3. Characteristics of the datasets containing general and dental variables

Characteristics of the datasets containing general variables

Prevalence of prediabetes in the sample containing general predictors (N = 6346) was 23.43%. A majority considered that they were not at risk of diabetes (N = 4678), were female (N = 3340), belonged to a race other than (non-Hispanic) White (N = 3839), were US citizen (N = 5578), were married or living with partner (N = 4627), used alcohol (defined as having had at least 12 drinks of any type of alcoholic beverage in any year; N = 4429), did not use any tobacco product last 5 days (N = 4921), were not diagnosed as hypertensive (N = 4463), had no self-reported hepatitis B (N = 6289) or hepatitis C (N = 6280), were not diagnosed jaundiced (N = 6217), had

no familial diabetes (N = 3925), were hepatitis B surface antibody negative (N = 4571), were hepatitis E IgG negative (N = 6061), had an education level of college/ Associate of Arts (AA) degree/above (N = 2915), did not engage in vigorous activity (N = 3710) but had moderate activity (N = 4493), had no gestational diabetes (only 172 females reported of gestational diabetes), had no overweight baby at birth (only 327 females reported of having had an overweight baby at birth), had no hysterectomy (only 525 females reported of having had hysterectomy), had no bilateral ovariectomy (only 281 females reported of having had bilateral ovariectomy), and had no female hormone intake (only 482 females reported of having had female hormones).

The sample had a mean age of 40.68 years (SD = 20.45 years; range=12-80 years), a mean income poverty ratio of 2.44 on a continuous scale ranging 0-5 (SD = 1.65), mean food security of 3.40 (SD = 0.97) on a continuous scale ranging 1-4, a mean daily TV watching duration of 2.34 hours (SD = 1.63 hours), a mean body mass index (BMI) of 27.73 kg/m^2 (SD = 7.09 kg/m^2), a mean waist circumference of 94.59 cm (SD = 17.40 cm), a mean white blood cell (WBC) count of 7.22×10^9/L (SD = 2.27×10^9/L), a mean monocyte count of 0.58 $\times10^9$/L (SD = 0.20×10^9/L), a mean red blood cell (RBC) count of 4.68 million cells/uL (SD = 0.49 million cells/uL), a mean hemoglobin level of 13.97 g/dL (SD = 1.49 g/dL), a mean alanine amino transferase (ALT) level of 23.63 IU/L (SD = 18.20 IU/L), a mean aspartate amino transferase (AMT) level of 24.87 IU/L (SD = 17.54 IU/L), a mean serum total calcium level of 9.49 mg/dL (SD = 0.36 mg/dL), a mean serum globulin level of 2.81 g/dL (SD = 0.43 g/dL), a mean gamma glutamyl transferase (GGT) level of 24.34 U/L (SD = 33.51 U/L), a mean serum iron level of 84.16 ug/dL (SD = 36.98 ug/dL), a mean serum potassium level of 4.02 mmol/L (SD = 0.35 mmol/L), a mean osmolality of 278.84 mmol/kg (SD = 4.71 mmol/kg), a mean serum phosphorus level of 3.95 mg/dL (SD = 0.65 mg/dL), a mean serum triglyceride level of 134.27 mg/dL (SD = 97.56 mg/dL), a mean serum uric acid level of 5.31 mg/dL (SD = 1.38 mg/dL), a mean systolic blood pressure (SBP) of 119.32 mm Hg (SD = 17.28 mm Hg), a mean diastolic blood pressure (DBP) of 66.85 mm Hg (SD = 13.09 mm Hg), and a mean hematocrit of 41.34 (SD = 4.04).

Characteristics of the datasets containing dental variables

Prevalence of prediabetes in the sample containing dental predictors (N = 3167) was 29.46%. A majority were US citizen (N = 2728), female (N = 1668), married or living with partner (N =

2260), belonged to a race other than (non-Hispanic) White (N = 1762), had never undergone periodontal therapy (N=2409), had no periodontitis (N = 1968), had no self-reported tooth mobility (N = 2708), and had an education of college/AA degree/college graduate or above (N = 1897). The sample had a mean age of 51.04 years (SD = 14.24 years; range: 30-80 years), a mean income poverty ratio of 2.69 on a continuous scale ranging 0-5 (SD = 1.67), and a mean number of teeth of 24.34 (SD = 6.41).

Univariate distributional analyses

Distribution of the characteristics of the NHANES 2013-2014 data samples prepared for building prediction models using general (N =6346) and dental (N = 3167) attributes between prediabetic and non-prediabetic individuals are summarized in **table 1.1** and **table 1.2**, respectively. According to the univariate distributional analyses of the categorical variables in the dataset containing general predictors, self-perceived DM risk, males, diagnosed hypertension, presence of hepatitis E IgG, college/AA degree/above education level, moderate activity, having an overweight baby at birth, hysterectomy, bilateral ovariectomy, and female hormone intake were significantly higher in the prediabetic group (N = 1487) while the presence of hepatitis B surface antibody, education level of 9-11 grade, and vigorous activity were significantly higher in the non-prediabetic group (N = 4859). Among continuous variables, mean values of age, duration of watching TV, BMI, waist circumference, RBC count, hemoglobin, ALT, AMT, serum calcium, serum globulin, GGT, osmolality, serum uric acid, mean SBP, mean DBP, and hematocrit were significantly higher in the prediabetic group while food security, serum potassium and serum phosphorus were significantly higher in the non-prediabetic group (**table 1.1**).

Corresponding univariate analyses of the categorical variables in the dataset consisting of dental variables revealed that periodontitis, males, self-reported tooth mobility, education <9th grade and 9-11 grade were significantly higher in the prediabetic group (N = 933) while college/AA degree/college graduate/above education level was significantly higher in the non-prediabetic group (N = 2234). Among continuous variables, mean age was significantly higher in the prediabetic group while the mean values of income-poverty ratio and the number of teeth were significantly higher in the non-prediabetic group (**table 1.2**).

3.4. Features selected for modelling

Feature selection algorithms employed upon the train dataset containing the 156 general attributes and the features extracted by each method are given in the **table 2**. A descriptive summary of the ultimate set of 46 variables selected for modelling that was decided upon considering both the output of the feature selection algorithms and the evidence from the comprehensive literature review of the determinants of prediabetes (**appendix: table 3**) is given in the **table 1.1**. This comprised of 22 categorical variables, namely, self-perceived DM risk, gender, race, citizenship, marital status, alcohol use, past any tobacco use, diagnosed hypertension, hepatitis B, hepatitis C, diagnosed jaundice, familial diabetes, hepatitis B surface antibody, hepatitis E IgG, education, vigorous activity, moderate activity, gestational DM, overweight baby at birth, hysterectomy, bilateral ovariectomy, and female hormone intake, and 24 numeric variables, namely, age, income-poverty ratio, food security, duration of watching TV, BMI, waist circumference, WBC count, monocyte count, RBC count, hemoglobin level, serum ALT, serum AMT, serum calcium, serum globulin, serum GGT, serum iron, serum potassium, osmolality, serum phosphorus, triglyceride level, serum uric acid, mean SBP, mean DBP, and hematocrit.

3.5. Optimal models and elucidated determinants

Twenty determinants of prediabetes in total encompassing various socio-economic, physiological, and biochemical variables, namely, age, income-poverty ratio, marital status, food security, citizenship, mean SBP, RBC count, serum triglyceride level, hematocrit, serum GGT, serum uric acid, diagnosed hypertension, hepatitis C, ALT, osmolality, serum potassium, vigorous activity, monocyte count, serum calcium, and hysterectomy, were elucidated by one or more of the 3 optimal logistic regression models as shown in the **table 3.1**. The 3 optimal models were the logistic regression with original, un-resampled train data, with majority class under-sampling and with minority-class oversampling which had AUC of 70.76%, 70.30% and 70.83% respectively, upon the internal validation dataset.

Four optimal models were produced by non-linear/ensemble machine learning algorithms, namely, RF with minority class oversampling, RF with SMOTE, ANN with original, un-resampled data and GB with original, un-resampled data which had AUC of 71.59%, 70.66%, 70.21% and 70.55% respectively, upon the internal validation dataset (N = 3172) (**table 3.2**). The optimal ANN model built from the original, un-resampled data via a linear output function

was a feed-forward, five-fold cross-validated neural network containing uncorrelated (correlation coefficient <0.75), automatically standardized variables with tuned parameters of one hidden layer, decay parameter of 0.1, 24 nodes in the hidden layer, 5 neural networks trained with different random number seeds and their predictions averaged. The other 3 optimal models were ten-fold cross-validated ensembles containing automatically standardized variables with default functions and parameters. Via one or more of these 4 optimal non-linear/ensemble models, 25 variables were elucidated as important predictors of prediabetes, which consisted of the same 20 predictors elicited by logistic regression models and 5 additional predictors, namely, waist circumference, BMI, WBC count, hepatitis B and AMT (**table 3.2**). Two predictors were common to all 7 optimal (3 logistic regression and 4 non-linear/ensemble) models, namely, age and serum potassium. None of the bagged CART models reached the benchmark AUC of 70% and hence were not presented.

None of the 25 models built using dental variables reached the benchmark of 70% AUC and therefore the best models having the highest AUC built using each algorithm and the elucidated predictors are presented in the **table 4**. The AUC (range: 54.53%-59.43%) estimates were sub-optimal. The optimal ANN model built from the ROSE resampled data (AUC on the internal validation data = 58.21%) via a linear output function was a feed-forward, five-fold cross-validated neural network containing automatically standardized variables with tuned parameters of one hidden layer, decay parameter of 0.1, 7 nodes in the hidden layer, 5 neural networks trained with different random number seeds and their predictions averaged. The other optimal models built using RF with ROSE resampling, GB with ROSE resampling and bagged CART with original data were ten-fold cross-validated ensembles containing automatically standardized variables with default functions and parameters. The logistic regression model with ROSE resampling (AUC on the internal validation data = 59.43%) was the best performing model and elucidated 5 predictors including periodontitis and the self-reported tooth mobility. Periodontitis was an important determinant of prediabetes as per all the other 4 non-linear/ensemble models as well (**table 4**).

3.6. Benchmarking

As shown in the **table 5**, the optimal model built on the train data consisting of the general predictors (N = 3174) outperformed the CDC prediabetes screening tool on both internal (N =

3172) and external (N = 3000) validation data with 71.59% and 70.01% AUC respectively. The corresponding AUC of the CDC prediabetes screening tool on the internal (N =3172) and external (N = 3000) validation data were 64.40% and 62.80%, respectively. As per the statistical test for comparing two ROC curves by Hanley & McNeil (1982), AUC of the optimal model was significantly higher than that of the CDC prediabetes screening tool on both internal (absolute difference = 7.19%, z = 4.3086, non-directional/two-tailed p value = 0.000017) and external (absolute difference = 7.21%, z = 4.2227, non-directional/two-tailed p value = 0.000024) validation data.

The performance of the optimal model built on the train data consisting of the dental predictors (N = 1584) was comparable to that of the CDC prediabetes screening tool on both internal (N = 1583) and external (N = 1500) validation data which demonstrated 59.43% and 57.88% AUC respectively. The corresponding AUC of the CDC prediabetes screening tool on the internal (N =1583) and external (N = 1500) validation data were 59.10% and 58.80%, respectively (**table 5**). As per the statistical test for comparing two ROC curves by Hanley & McNeil (1982), there was no significant difference between AUC of the optimal model and that of the CDC prediabetes screening tool on both internal (absolute difference = 0.33%, z = 0.146, non-directional/two-tailed p value = 0.883922) and external (absolute difference = 0.92%, z = 0.3957, non-directional/two-tailed p value = 0.692326) validation data.

4. DISCUSSION

A number of socio-economic, physiological and biochemical determinants of prediabetes were elucidated by machine learning models having an optimal predictive power. The set of predictors included a few new potential attributes as well. Models built using dental variables were comparatively less predictive of prediabetes where periodontitis and self-reported tooth mobility, which is a proxy indicator of periodontal disease appeared significant dental determinants of the disease. Models built using general predictors outperformed while those containing dental attributes were on a par with the chosen benchmark, i.e. the CDC prediabetes screening tool.

4.1. Socio-economic determinants

Age, income-poverty ratio, marital status, food security, and citizenship were predictors of prediabetes elucidated by both linear and non-linear machine learning models; age and not being a US citizen increased the odds of having prediabetes while higher income-poverty ratio, being married, and higher food security reduced the odds of being prediabetic as per logistic regression models giving valuable insights into the impact of socio-economic factors on the disease. In fact, age is the single most important predictor which has been allocated the highest score in both the ADA prediabetes screening test (ADA prediabetes screening test, 2018) and the CDC prediabetes risk prediction tool (CDC prediabetes screening test, 2018). As suggested by previous studies, the co-existence of socio-economic risk factors such as advancing age, lower income, diminished food security, unmarried status and not having the citizenship may create a socio-economic gradient of the disease prevalence favoring its aggregation among the socio-economically-deprived groups (Meranus, 2015; Cai et al., 2017). Therefore, public health programs to curb prediabetes among the US population can be rendered more effective by addressing these socio-economic determinants which tend to place disadvantaged groups at higher risk.

4.2. Clinical, physiological, and biochemical determinants

4.2.1. Anthropometry

Multiple anthropometric measures are associated with overt diabetes (Borné et al., 2015). Waist circumference and BMI were predictors of prediabetes elucidated by non-linear/ensemble

models in the present study and were consistent with the findings of Shearer et al. (2016) and Tao et al. (2017) who reported that anthropometric attributes were proven markers of future dysglycemic risk and might act as potentially useful variables for designing an individualized approach to early diagnostic and preventive interventions along the disease trajectory. A plethora of research supports the predictive value of anthropometry in prediabetes and diabetes risk modelling and anthropometric attributes are also included in both the ADA prediabetes screening test (ADA prediabetes screening test, 2018) and the CDC prediabetes risk score tool (CDC prediabetes screening test, 2018).

4.2.2. Vigorous physical activity

It is well-known that physical activity exerts a sustained protective effect against the onset of prediabetes (Chow et al., 2016). However, vigorous physical activity was found to have a protective effect in the present study while moderate physical activity was not predictive of prediabetes. Findings are congruous with Jung et al. (2015) who revealed that high-intensity interval training (HIIT) is more effective than moderate-intensity continuous training (MICT) for prediabetic adults and Earnest (2008) who proposed that interval training – a type of vigorous exercise improves physiology of prediabetic individuals. Reversal of vascular endothelium-dependent dysfunction (Liu et al., 2013), maintenance of a proper lean-to-fat balance with respect to body mass, increase in the insulin sensitivity and glucose tolerance (Helmrich et al., 1991) have been suggested as the possible mechanisms by which physical activity exerts a protective effect on prediabetes.

4.2.3. Triglycerides

Elevated triglyceride levels are a well-established predictor of prediabetes (Deepa et al., 2015; Akehi et al., 2010) and may even be useful as a risk profiling tool as Abbasi et al. (2016) reported that hypertriglyceridemia can be used as a simple approach to identify insulin resistance and enhanced cardio-metabolic risk in patients with prediabetes. There is evidence to support the premise that alterations of the plasma lipidome including hypertriglyceridemia observed in overt diabetes may in fact prevail in prediabetes as well (Meikle et al., 2013), further warranting interventions aimed at early dysglycemic states. Since insulin resistance and obesity are associated with adipose tissue inflammation and increased production of coagulation factors

(Lallukka et al., 2017), triglyceridemia in prediabetic individuals is likely to aggravate their negative hemodynamic profile and elevate the risk of adverse cardiovascular outcomes.

4.2.4. Hysterectomy

There is some evidence to support that hysterectomy is associated with diabetes (Luo et al., 2017; Wilson & Mishra,2017; Appiah et al., 2014) but little evidence exists for its possible association with prediabetes. Noteworthily, high levels of evidence from clinical trials are available that postmenopausal therapy with estrogen alone in women who have had a hysterectomy may reduce the incidence of diabetes (Bonds et al., 2006). Experimental data confirmed a protective role exerted by estrogens on glucose metabolism (Le May et al., 2006) while women who had undergone hysterectomy experienced a sudden, marked reduction in their estrogen production (Korse et al., 2009) which might be a plausible explanation for their elevated risk of diabetes. The present study indicates that this association might be manifest at the prediabetic phase as well.

4.2.5. Hemodynamics

Mean SBP, osmolality and being diagnosed hypertensive were positively associated with prediabetes in this study reflecting the role of hemodynamic predictors in its pathogenesis. Since hypertension is an established risk factor for prediabetes (Casapulla et al., 2017; Okwechime et al., 2015) manifest as impaired fasting glucose, impaired glucose tolerance, and the metabolic syndrome where the last condition is characterized by hypertension along with several other metabolic disorders, Garber (2011) argued that blood pressure control strategies in prediabetes ought to be similar to those applied in overt diabetes. Though it is well-known that blood viscosity is elevated in diabetes due to the rise in osmolality causing increased capillary permeability leading to a rise in hematocrit, Irace et al. (2014) was the first to report a direct relationship between blood viscosity and blood glucose suggesting that the association is observable in normoglycemic and diabetic precursor stages as well. This supports the hypothesis that alterations are very early and observable even in physiologically normal states. Therefore, hemodynamic profile, especially osmolality, can be a potential source for early identification of high-risk individuals before the onset of prediabetes.

4.2.6. Hematological markers

Increased hematocrit and red cell count and reduced monocyte count were associated with prediabetes which are consistent with findings from previous studies. For instance, an increased RBC count, partially attributable to elevated HbA1c levels (Simmons; 2010) as well as elevated hematocrit possibly due to increased blood viscosity and consequent rise in capillary permeability (Meisinger et al., 2014) were found to be associated with prediabetes. Noteworthily, WBC count was also found an important predictor exclusively by ensemble models in the present study, namely RF and GB, indicating a complex, non-linear relationship with the disease. A rise in WBC count was found in incident diabetes (Twig et al., 2013; Vozarova et al., 2002), whereas a meta-analysis (GKrania-Klotsas et al., 2010) provided further insights by revealing that elevations in total WBC, constituent neutrophil and lymphocyte counts but reductions in monocyte count were associated with diabetes. As Twig et al. (2013) reported that WBC count is an independent risk factor for diabetes even within the normal range, it is possible that changes manifest much earlier in the natural course of the disease. Observed changes in the number of white cells suggest that a chronic activation of the immune system may play a role in the pathogenesis and further research may hence be warranted to explore the role of different white cell types in the early course of the disease.

4.2.7. Liver function profile

Although hepatitis C virus (HCV) infection's association with both prediabetes (Burman et al., 2015; Ali et al., 2014; Mukhtar et al., 2012) and diabetes (Howard et al., 2003; Wang et al., 2003) is well established and possible hypotheses for its pathogenesis such as HCV infection-induced hepato-steatosis leading to metabolic changes that result in diabetes (Alexander, 2000) and HCV-infection induced disruption of insulin signaling pathways triggering abnormalities in glucose homeostasis (Mukhtar et al., 2012) have been put forward, the impact of hepatitis B virus (HBV) infection on prediabetes is not as straightforward. Despite having some evidence for an association of HBV infection with overt diabetes (Gisi et al., 2017; Khalili et al., 2015), there is little evidence for its effect on precursor stages of diabetes. However, the present study found via non-linear machine learning models that HBV is a predictor of prediabetes. Conversely, elevated ALT and GGT levels are consistently associated with both prediabetes (Fei et al., 2012; Nguyen et al., 2011) and diabetes (Ko et al., 2015; Kubo et al., 2007), but the relationship of AMT levels with either condition is not clear although Al-Jameil et al. (2014) found a weak

association between AMT and T2DM. The present study found AMT was a predictor of prediabetes as elucidated by non-linear models indicating a possibly non-linear association. On the whole, the inflammatory milieu associated with chronic viral infections as well as liver damage indicated by elevated liver enzymes stand as strong predictors of prediabetes as shown by our study.

4.2.8. Molecular serum markers

Serum uric acid, potassium and calcium levels were found to be predictors of prediabetes in the present study. Associations of elevated serum uric acid (Anothaisintawee et al., 2017; Vučak et al., 2012; Chu et al., 2017) and low serum potassium (Meisinger et al., 2013) with prediabetes have been reported. The causal mechanism behind elevated serum uric acid in prediabetes remains largely inconclusive although a possible negative impact on pancreatic β-cell function has been suggested (van der Schaft et al., 2017). Similarly, the pathophysiological basis of the association of hypokalemia with prediabetes is also currently inconclusive although Meisinger et al. (2013) who first revealed this association argued that it could be due to decreased insulin secretion or a higher ratio of proinsulin to insulin secretion triggered by low serum potassium levels. Despite a lack of evidence for the association of elevated serum calcium with prediabetes, a number of studies reported its association with overt diabetes (Rooney et al., 2016; Suh et al., 2017; Sing et al., 2016; Fu et al., 2015). It is possible that aberrations of calcium homeostasis are indeed manifest earlier in the natural history of the disease than suggested by previous studies. Rooney et al. (2016) hypothesized that altered calcium homeostasis causes abnormal β cell functioning leading to dysglycemia which could be the pathological mechanism underlying this association.

4.3. Potentially novel determinants

It is noteworthy that several established biomarkers of elevated risk of diabetes albeit with only sparse evidence for their associations with precursor stages of the disease were identified as potential determinants of prediabetes in the present study. Markers of the early stages of the disease may help timely diagnosis of high-risk individuals prior to developing diabetes, where standard risk factors may not corroborate an early diagnosis. According to Suvitaival et al. (2018), such markers may be present years before the onset of diabetes, which a machine-

intelligence based approach may identify unlike by a linear model. Most of the markers of prediabetes spanning socio-economy (age, income-poverty ratio, marital status, food security, and citizenship), anthropometry (waist circumference, BMI), hemodynamics (mean SBP, osmolality, and diagnosed hypertension), life-style (vigorous activity), lipidome (serum triglycerides), hematology (RBC count, WBC count, hematocrit and monocyte count), liver function profile (GGT, ALT, and hepatitis C) and serum biomarkers (uric acid & potassium) identified in the present study are established determinants of the disease and are reinforced by a substantial body of evidence as discussed above. However, several known diabetes risk markers newly identified as determinants of prediabetes in the present study such as serum calcium, hysterectomy, hepatitis B and AMT (Abbasi et al., 2012) provide directions for future research which could be potential indicators of the precursor stages of diabetes.

4.4. Dental determinants

While prediabetes screening in a dental clinic setting has been proven effective (AlGhamdi et al., 2013; Maples et al., 2015), public health databases might not yield high predictive power when dental predictors are used in stand-alone models as revealed by the present study. However, further research is required to assess if a combined model containing both dental and general predictors from a public health database could substantially enhance predictive power. Moreover, since the models with dental predictors were built using a smaller train sample and a narrower age range, a larger sample size having a wider age range is suggested for future predictive modelling studies. Studies have found that periodontal disease is associated with dysglycemic states such as HbA1c progression (Demmer et al., 2010), incident diabetes (Demmer et al., 2008) and prediabetes (Arora et al., 2014; Mustapha, 2014; Ilievski et al., 2016) for which a possible microbial etiopathogenesis has been suggested (Demmer et al., 2015). The present study reinforced its value as a clinical marker of dysglycemia which can be used in a dental setting.

4.5. Strengths and limitations

Despite reports that prediabetes is more difficult to predict than diabetes (Choi et al., 2014), we built models with general attributes that outperformed the chosen benchmark, and the elucidated predictors are internally valid indicating their utility among the US population. Nevertheless,

their generalizability to non-US populations may be constrained by the cross-sectional study design and the context-specific nature of some variables. While the strategies such as using a mix of linear, non-linear and ensemble models, handling class imbalance via resampling methods, applying extensive feature selection methods and careful handling of missing data would have helped to construct models with general predictors that outperformed the CDC prediabetes screening tool, measures that were not used in the present study such as the use of a different set of algorithms and parameter tuning (Bergstra et al., 2013) might have the potential to further enhance their predictive power and hence are suggested to be implemented in future studies.

To the best of our knowledge this is the first study that applied an extensive methodology comprised of a range of feature selection methods and machine learning algorithms on a single NHANES database to elucidate the determinants of prediabetes. As recommended by Collins et al. (2011), a systematic approach was adopted in the present study to select attributes, apply algorithms, and handle missing data which enabled to not only produce models with adequate predictive power also identify a few novel predictors of prediabetes which may provide avenues for emerging areas of prediabetes research. Many well-established determinants were also elucidated standing as proof of concept for our analytical approach. For instance, all 20 determinants elucidated by logistic regression models were identified as important predictors by other non-linear machine learning models while the latter also identified 5 additional predictors. This is comparable to the big-data based, hypothesis-free machine learning method endorsed by Anderson et al. (2016) who developed powerful prediction models for progression to diabetes using high-dimensional electronic health records. The present study thus demonstrates that a machine intelligence-based approach is highly applicable to studies that use public health databases to elucidate determinants of prediabetes as it has the potential to generate new knowledge and trajectories for future research. While inherent pitfalls of a cross-sectional study design may have affected the present study, it is noteworthy that the constructed prediction models contained 46 independent variables and were adjusted for many potential confounders enhancing the validity of findings.

An inherent limitation of non-linear and ensemble machine learning algorithms is the diminished interpretability; directionality of associations cannot be easily illustrated as via a linear model such as logistic regression (Cafri et al., 2016). While data mining by non-linear and ensemble algorithms offer superior predictive performance than conventional parametric models,

deciphering variable effects may prove difficult. Therefore, novel predictors identified by such algorithms should be evaluated in conjunction with the current knowledge and existing body of evidence while further research is recommended to elucidate the pathophysiology underlying those non-linear, complex associations with prediabetes. Since we used data from a national, cross-sectional study, elucidated associations do not indicate temporality and further studies, preferably prospective cohort studies, are required to elicit directionality especially in relation to novel predictors. In addition, cross-sectional studies are known to suffer from a range of biases and confounding although the latter effect may have been greatly reduced due to the advanced feature selection tools applied and multivariate models consisting of many potential confounders that were constructed. However, causality is not indicated in these cross-sectional associations owing to the limitations of the study design.

4.6. Methodological issues

An empirical issue is posed by the differences in the definitions of prediabetes used in various settings. In addition, the use of self-reported data on ever-being diagnosed as having prediabetes may not be prudent for the purpose of outcome classification, because, unlike diabetes, prediabetes is often undiagnosed due to its asymptomatic nature (Bansal, 2015), is reversible so that some self-reported ever-prediabetic individuals may have shifted to normoglycemic status or alternatively progressed to diabetic stage along the disease spectrum when the data were collected. In fact, it has been reported that up to 70% of individuals with prediabetes will eventually develop diabetes, with around 5–10% of people with prediabetes becoming diabetic annually (Tabá et al., 2012). The feasibility of excluding self-reported data on prediabetes was reinforced by the ROC analysis performed in the present study which demonstrated their poor discriminant power (**figure 1**).

Missing data management was reported as a critical area that lacked quality in a systematic review of the methodological issues of prediction models of prediabetes (Collins et al., 2011). In a predictive modelling study by Ogunyemi et al. (2015) the cut-off for exclusion of variables was set at 50% of missing data, but, in the present study, variables having >30% missing data were excluded and thus, despite having a much larger sample size, a more conservative approach was adopted. An extensive, post-hoc accuracy analysis of the multiple imputation was also conducted

which was proved to be adequate with only a few variables having significantly different distributions between original and imputed samples.

The model performance is context-dependent; a prediction model based on a clinical database having a higher prevalence of the condition of interest usually achieves a high AUC whereas models built on imbalanced, public health data which closely resemble the true and essentially lower prevalence of a condition may not achieve comparable AUC estimates. For example, a prospective, population-based cohort study (N = 10,038) on diabetes screening demonstrated AUC values of 70% and 69% for HbA1C and FPG tests, respectively (Choi et al., 2011).

4.7. Public health implications

Although the use of biomarkers in prediction models increases predictive value (Abbasi et al., 2013), their applicability in community screening tools is fraught with difficulty because large-scale collection of some biological data such as blood specimens may be imprudent due to high cost, required technical expertise, and ethical constraints. Also, special care may be required to achieve safe storage of biological specimens collected in public health surveys in the context of challenges like bio-terrorism (Regidor, 2004). However, simpler markers such as directly measurable anthropometry or non-invasively measurable saliva- or urine- based markers may be useful for community-level diagnosis purposes. As conventional prediabetes screening tools reportedly fail to diagnose a large proportion of undetected prediabetic individuals (Dall et al., 2014), these simpler biomarkers may be useful as additional predictors to enhance the predictive ability of screening tools.

The broad range of socio-economic, physiological, and biochemical predictors elucidated by the present study may be applied for risk profiling to effectively capture the individuals who are most likely to develop prediabetes. Some of the predictors may not be efficient for use on a population level though they may be useful in clinical settings. Therefore, further research is required to consolidate the determinants elucidated by the present study which may eventually be used in clinical and community settings.

4.8. Conclusions

In conclusion, the present study elucidated a range of socio-economic, physiological, and biochemical determinants of prediabetes including a few potentially novel associations via optimal prediction models with adequate predictive power. Dental models by themselves were not predictive of prediabetes although periodontal disease was associated with the disease. Models containing general predictors outperformed whilst those with dental predictors equaled the chosen benchmark i.e., the CDC prediabetes screening tool. The present study, thus, demonstrated the potential applicability of a systematic, machine intelligence-based modelling approach on a public health database to elucidate the determinants of prediabetes and provided novel insights into its etio-pathogenesis.

REFERENCES

Abbasi, A., 2013. *Biomarkers and Prediction Model for Type 2 Diabetes and Diabetes Related Outcomes*. University of Groningen Library][Host].

Abbasi, A., Bakker, S.J., Corpeleijn, E., Gansevoort, R.T., Gans, R.O., Peelen, L.M., van der Schouw, Y.T., Stolk, R.P., Navis, G., Spijkerman, A.M. and Beulens, J.W., 2012. Liver function tests and risk prediction of incident type 2 diabetes: evaluation in two independent cohorts. *PloS one*, *7*(12), p.e51496.

Abbasi, F., Kohli, P., Reaven, G.M. and Knowles, J.W., 2016. Hypertriglyceridemia: A simple approach to identify insulin resistance and enhanced cardio-metabolic risk in patients with prediabetes. *diabetes research and clinical practice*, *120*, pp.156-161.

ADA prediabetes screening test, 2018. Available online at http://main.diabetes.org/dorg/PDFs/risk-test-paper-version.pdf

Aekplakorn, W., Tantayotai, V., Numsangkul, S., Sripho, W., Tatsato, N., Burapasiriwat, T., Pipatsart, R., Sansom, P., Luckanajantachote, P., Chawarokorn, P. and Thanonghan, A., 2015. Detecting prediabetes and diabetes: agreement between fasting plasma glucose and oral glucose tolerance test in Thai adults. *Journal of diabetes research*, *2015*.

Alexander, G.J., 2000. An association between hepatitis C virus infection and type 2 diabetes mellitus: What is the connection?. *Annals of internal medicine*, *133*(8), pp.650-652.

AlGhamdi, A.S.T., Bukhari, S.M., Elias, W.Y., Merdad, K. and Sonbul, H., 2013. Dental clinics as potent sources for screening undiagnosed diabetes and prediabetes. *The American journal of the medical sciences*, *345*(4), pp.331-334.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S., 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PloS one*, *12*(7), p.e0179805.

American Diabetes Association, 2014. Standards of Medical Care in Diabetes—2014. Diabetes Care 2014; 37 (Suppl. 1): S14–S80Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 2014; 37 (Suppl. 1): S81–S90. *Diabetes care*, *37*(3), pp.887-887.

Anderson, J.P., Parikh, J.R., Shenfeld, D.K., Ivanov, V., Marks, C., Church, B.W., Laramie, J.M., Mardekian, J., Piper, B.A., Willke, R.J. and Rublee, D.A., 2016. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*, *10*(1), pp.6-18.

Andriankaja, O.M. and Joshipura, K., 2014. Potential association between prediabetic conditions and gingival and/or periodontal inflammation. *Journal of diabetes investigation*, *5*(1), pp.108-114.

Arora, N., Papapanou, P.N., Rosenbaum, M., Jacobs, D.R., Desvarieux, M. and Demmer, R.T., 2014. Periodontal infection, impaired fasting glucose and impaired glucose tolerance: results from the Continuous National Health and Nutrition Examination Survey 2009–2010. *Journal of clinical periodontology*, *41*(7), pp.643-652.

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), pp.40-49.

Bansal, N., 2015. Prediabetes diagnosis and treatment: A review. *World journal of diabetes*, *6*(2), p.296.

Barber, S.R., Davies, M.J., Khunti, K. and Gray, L.J., 2014. Risk assessment tools for detecting those with pre-diabetes: a systematic review. *Diabetes research and clinical practice*, *105*(1), pp.1-13.

Bekkar, M., Djemaa, H.K. and Alitouche, T.A., 2013. Evaluation measures for models assessment over imbalanced datasets. *J Inf Eng Appl*, *3*(10).

Bergstra, J., Yamins, D. and Cox, D., 2013, February. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning* (pp. 115-123).

Bonds, D.E., Lasser, N., Qi, L., Brzyski, R., Caan, B., Heiss, G., Limacher, M.C., Liu, J.H., Mason, E., Oberman, A. and O'sullivan, M.J., 2006. The effect of conjugated equine oestrogen on diabetes incidence: the Women's Health Initiative randomised trial. *Diabetologia*, *49*(3), pp.459-468.

Borné, Y., Nilsson, P.M., Melander, O., Hedblad, B. and Engström, G., 2015. Multiple anthropometric measures in relation to incidence of diabetes: a Swedish population-based cohort study. *The European Journal of Public Health*, *25*(6), pp.1100-1105.

Breiman, L., 1996. Bagging predictors. *Machine learning*, *24*(2), pp.123-140.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Buuren, S.V. and Groothuis-Oudshoorn, K., 2010. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, pp.1-68.

Cafri, G. and Bailey, B.A., 2016. Understanding variable effects from black box prediction: Quantifying effects in tree ensembles using partial dependence. *Journal of Data Science*, *14*(1), pp.67-95.

Cai, L., Li, X., Cui, W., You, D. and Golden, A.R., 2017. Trends in diabetes and pre-diabetes prevalence and diabetes awareness, treatment and control across socioeconomic gradients in rural southwest China. *Journal of Public Health*, pp.1-6.

CDC prediabetes screening test, 2018. Available online at https://www.cdc.gov/diabetes/prevention/pdf/prediabetestest.pdf

Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 875-886). Springer, Boston, MA.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.

Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.H., Kang, E.S. and Kim, D.W., 2014. Screening for prediabetes using machine learning models. *Computational and mathematical methods in medicine, 2014*.

Choi, S.H., Kim, T.H., Lim, S., Park, K.S., Jang, H.C. and Cho, N.H., 2011. Hemoglobin A1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction: a 6-year community-based prospective study. *Diabetes care*, *34*(4), pp.944-949.

Chow, L.S., Odegaard, A.O., Bosch, T.A., Bantle, A.E., Wang, Q., Hughes, J., Carnethon, M., Ingram, K.H., Durant, N., Lewis, C.E. and Ryder, J., 2016. Twenty year fitness trends in young adults and incidence of prediabetes and diabetes: the CARDIA study. *Diabetologia*, *59*(8), pp.1659-1665.

Collins, G.S., Mallett, S., Omar, O. and Yu, L.M., 2011. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, *9*(1), p.103.

Dall, T.M., Narayan, K.V., Gillespie, K.B., Gallo, P.D., Blanchard, T.D., Solcan, M., O'Grady, M. and Quick, W.W., 2014. Detecting type 2 diabetes and prediabetes among asymptomatic adults in the United States: modeling American Diabetes Association versus US Preventive Services Task Force diabetes screening guidelines. *Population health metrics*, *12*(1), p.12.

Demmer, R.T., Jacobs, D.R. and Desvarieux, M., 2008. Periodontal disease and incident type 2 diabetes: results from the First National Health and Nutrition Examination Survey and its epidemiologic follow-up study. *Diabetes care*, *31*(7), pp.1373-1379.

Demmer, R.T., Desvarieux, M., Holtfreter, B., Jacobs, D.R., Wallaschofski, H., Nauck, M., Völzke, H. and Kocher, T., 2010. Periodontal status and A1C change: longitudinal results from the study of health in Pomerania (SHIP). *Diabetes care*, *33*(5), pp.1037-1043.

Demmer, R.T., Jacobs Jr, D.R., Singh, R., Zuk, A., Rosenbaum, M., Papapanou, P.N. and Desvarieux, M., 2015. Periodontal bacteria and prediabetes prevalence in ORIGINS: the oral infections, glucose intolerance, and insulin resistance study. *Journal of dental research*, *94*(9_suppl), pp.201S-211S.

Dong, X.L., Liu, Y., Sun, Y., Sun, C., Fu, F.M., Wang, S.L. and Chen, L., 2011. Comparison of HbA1c and OGTT criteria to diagnose diabetes among Chinese. *Experimental and clinical endocrinology & diabetes*, *119*(06), pp.366-369.

Echouffo-Tcheugui, J.B., Narayan, K.M., Weisman, D., Golden, S.H. and Jaar, B.G., 2016. Association between prediabetes and risk of chronic kidney disease: a systematic review and meta-analysis. *Diabetic Medicine*, *33*(12), pp.1615-1624.

Edwards, C.M. and Cusi, K., 2016. Prediabetes. *Endocrinology and Metabolism Clinics*, *45*(4), pp.751-764.

Fonti, V. and Belitser, E., 2017. Feature Selection using LASSO.

Friedman, J., Hastie, T. and Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), p.1.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), pp.367-378.

Fu, D.X., Cui, H.B., Guo, N.N., Su, N., Xu, J.X. and Wang, G.Y., 2016. Prediabetes and the risk of pancreatic cancer: a meta-analysis. *International Journal of Clinical and Experimental Medicine*, *9*(10), pp.19474-19479.

Garber, A.J., 2011. Hypertension and lipid management in prediabetic states. *The Journal of Clinical Hypertension*, *13*(4), pp.270-274.

Gevrey, M., Dimopoulos, I. and Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), pp.249-264.

Ghazanfari, Z., Haghdoost, A.A., Alizadeh, S.M., Atapour, J. and Zolala, F., 2010. A comparison of HbA1c and fasting blood sugar tests in general population. *International journal of preventive medicine*, *1*(3), p.187.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), pp.308-319.

Grossi, S.G. and Genco, R.J., 1998. Periodontal disease and diabetes mellitus: a two-way relationship. *Annals of periodontology*, *3*(1), pp.51-61.

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

Han, L., Luo, S., Yu, J., Pan, L. and Chen, S., 2015. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE journal of biomedical and health informatics*, *19*(2), pp.728-734.

Hanley, J.A. and McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), pp.29-36.

Harris, E., 2002, January. Information Gain Versus Gain Ratio: A Study of Split Method Biases. In *ISAIM*.

Hellgren, M., Steiner, K.H. and Bennet, L., 2017. Haemoglobin A1c as a screening tool for type 2 diabetes and prediabetes in populations of Swedish and Middle-East ancestry. *Primary care diabetes*, *11*(4), pp.337-343.

Helmrich, S.P., Ragland, D.R., Leung, R.W. and Paffenbarger Jr, R.S., 1991. Physical activity and reduced occurrence of non-insulin-dependent diabetes mellitus. *New England journal of medicine*, *325*(3), pp.147-152.

Hilawe, E.H., Chiang, C., Yatsuya, H., Wang, C., Ikerdeu, E., Honjo, K., Mita, T., Cui, R., Hirakawa, Y., Madraisau, S. and Ngirmang, G., 2016. Prevalence and predictors of prediabetes and diabetes among adults in Palau: population-based national STEPS survey. *Nagoya journal of medical science*, *78*(4), p.475.

Holm, N.C.R., Belstrøm, D., Østergaard, J.A., Schou, S., Holmstrup, P. and Grauballe, M.B., 2016. Identification of Individuals With Undiagnosed Diabetes and Pre-Diabetes in a Danish Cohort Attending Dental Treatment. *Journal of periodontology*, *87*(4), pp.395-402.

Huang, Y., Cai, X., Mai, W., Li, M. and Hu, Y., 2016. Association between prediabetes and risk of cardiovascular disease and all-cause mortality: systematic review and meta-analysis. *bmj*, *355*, p.i5953.

Huang, Y., Cai, X., Qiu, M., Chen, P., Tang, H., Hu, Y. and Huang, Y., 2014. Prediabetes and the risk of cancer: a meta-analysis.

Jagelid, M. and Movin, M., 2017. A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning: Conversion prediction of Spotify Users-A Case Study.

Jaiswal, A., Tabassum, R., Podder, A., Ghosh, S., Tandon, N. and Bharadwaj, D., 2012. Elevated level of C-reactive protein is associated with risk of prediabetes in Indians. *Atherosclerosis*, *222*(2), pp.495-501.

Jayanthi, N., Babu, B.V. and Rao, N.S., 2017. Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, *4*(1), p.26.

Jiang, Y., Owei, I., Wan, J., Ebenibo, S. and Dagogo-Jack, S., 2016. Adiponectin levels predict prediabetes risk: the Pathobiology of Prediabetes in A Biracial Cohort (POP-ABC) study. *BMJ Open Diabetes Research and Care*, *4*(1), p.e000194.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*.

Kim, J.Y., Goran, M.I., Toledo-Corral, C.M., Weigensberg, M.J. and Shaibi, G.Q., 2015. Comparing glycemic indicators of prediabetes: a prospective study of obese Latino Youth. *Pediatric diabetes*, *16*(8), pp.640-643.

Korse, C.M., Bonfrer, J.M., Van Beurden, M., Verheijen, R.H. and Rookus, M.A., 2009. Estradiol and testosterone levels are lower after oophorectomy than after natural menopause. *Tumor Biology*, *30*(1), pp.37-42.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P., 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), pp.25-36.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. and Engelhardt, A., 2017. Caret: classification and regression training. 2016. *R package version*, *4*.

Kursa, M.B. and Rudnicki, W.R., 2010. Feature selection with the Boruta package. *J Stat Softw*, *36*(11), pp.1-13.

Lalla, E., Kunzel, C., Burkett, S., Cheng, B. and Lamster, I.B., 2011. Identification of unrecognized diabetes and pre-diabetes in a dental setting. *Journal of dental research*, *90*(7), pp.855-860.

Lalla, E. and Papapanou, P.N., 2011. Diabetes mellitus and periodontitis: a tale of two common interrelated diseases. *Nature Reviews Endocrinology*, *7*(12), p.738.

Lallukka, S., Luukkonen, P.K., Zhou, Y., Petäjä, E.M., Leivonen, M., Juuti, A., Hakkarainen, A., Orho-Melander, M., Lundbom, N., Olkkonen, V.M. and Lassila, R., 2017. Obesity/insulin resistance rather than liver fat increases coagulation factor activities and expression in humans. *Thrombosis and haemostasis*, *2017*(2), pp.207-428.

Lantz, B., 2013. *Machine learning with R*. Packt Publishing Ltd.

Lauron, M.L.C. and Pabico, J.P., 2016. Improved Sampling Techniques for Learning an Imbalanced Data Set. *arXiv preprint arXiv:1601.04756*.

Le May, C., Chu, K., Hu, M., Ortega, C.S., Simpson, E.R., Korach, K.S., Tsai, M.J. and Mauvais-Jarvis, F., 2006. Estrogens protect pancreatic β-cells from apoptosis and prevent insulin-deficient diabetes mellitus in mice. *Proceedings of the National Academy of Sciences*, *103*(24), pp.9232-9237.

Lee, M., Saver, J.L., Hong, K.S., Song, S., Chang, K.H. and Ovbiagele, B., 2012. Effect of pre-diabetes on future risk of stroke: meta-analysis. *BMJ*, *344*, p.e3564.

Lee, Y.H., Bang, H., Kim, H.C., Kim, H.M., Park, S.W. and Kim, D.J., 2012. A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores. *Diabetes care*, *35*(8), pp.1723-1730.

Lesmeister, C., 2017. *Mastering machine learning with r*. Packt Publishing Ltd.

Li, D.C., Liu, C.W. and Hu, S.C., 2010. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, *40*(5), pp.509-518.

Li, X., Liu, H., Du, X., Zhang, P., Hu, G., Xie, G., Guo, S., Xu, M. and Xie, X., 2016. Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 799). American Medical Informatics Association.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.

Löe, H., 1993. Periodontal disease: the sixth complication of diabetes mellitus. *Diabetes care*, *16*(1), pp.329-334.

Loh, W.Y., 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), pp.14-23.

Longadge, R. and Dongre, S., 2013. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.

Lunardon, N., Menardi, G. and Torelli, N., 2014. ROSE: A Package for Binary Imbalanced Learning. R Journal, 6(1).

Lusa, L., 2010. Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics*, *11*(1), p.523.

Mainous, A.G., Tanner, R.J., Baker, R., Zayas, C.E. and Harle, C.A., 2014. Prevalence of prediabetes in England from 2003 to 2011: population-based, cross-sectional study. *BMJ open*, *4*(6), p.e005002.

Maples, S., Aldasouqi, S., Little, R., Baughman, H., Joshi, M. and Salhi, R., 2015. Detection of Undiagnosed Prediabetes and Diabetes in Dental Patients: A Proposal of a Dental-Office-Friendly Diabetes Screening Tool. *Journal of Diabetes Mellitus*, *6*(01), p.25.

Marinov, M., Mosa, A.S.M., Yoo, I. and Boren, S.A., 2011. Data-mining technologies for diabetes: a systematic review. *Journal of diabetes science and technology*, *5*(6), pp.1549-1556.

Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A. and Tourassi, G.D., 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, *21*(2-3), pp.427-436.

Meikle, P.J., Wong, G., Barlow, C.K., Weir, J.M., Greeve, M.A., MacIntosh, G.L., Almasy, L., Comuzzie, A.G., Mahaney, M.C., Kowalczyk, A. and Haviv, I., 2013. Plasma lipid profiling shows similar associations with prediabetes and type 2 diabetes. *PloS one*, *8*(9), p.e74341.

Meranus, D.H., 2015. *The association between neighborhood socioeconomic status and the prevention and control of prediabetes and diabetes in King County, Washington* (Doctoral dissertation).

National Center for Health Statistics, 2018. *NCHS Research Ethics Review Board (ERB) Approval* Available at https://www.cdc.gov/nchs/nhanes/irba98.htm

Neumann, U., Genze, N. and Heider, D., 2017. EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData mining*, *10*(1), p.21.

Nowicka, P., Santoro, N., Liu, H., Lartaud, D., Shaw, M.M., Goldberg, R., Guandalini, C., Savoye, M., Rose, P. and Caprio, S., 2011. Utility of hemoglobin A1c for diagnosing prediabetes and diabetes in obese children and adolescents. *Diabetes care*, *34*(6), pp.1306-1311.

Ogunyemi, O. and Kermah, D., 2015. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 983). American Medical Informatics Association.

Okwechime, I.O. and Roberson, S., 2015. Prevalence and predictors of pre-diabetes and diabetes among adults 18 years or older in Florida: A multinomial logistic modeling approach. *PloS one*, *10*(12), p.e0145781.

Poltavskiy, E., Kim, D.J. and Bang, H., 2016. Comparison of screening scores for diabetes and prediabetes. *Diabetes research and clinical practice*, *118*, pp.146-153.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., 1996. *Numerical recipes in C* (Vol. 2). Cambridge: Cambridge university press.

Rahman, M.M. and Davis, D.N., 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, *3*(2), p.224.

Rathmann, W., Kowall, B., Heier, M., Herder, C., Holle, R., Thorand, B., Strassburger, K., Peters, A., Wichmann, H.E., Giani, G. and Meisinger, C., 2010. Prediction models for incident type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. *Diabetic medicine*, *27*(10), pp.1116-1123.

Regidor, E., 2004. The use of personal data from medical records and biological materials: ethical perspectives and the basis for legal restrictions in health research. *Social science & medicine*, 59(9), pp.1975-1984.

Romanski, P. and Kotthoff, L., 2009. Fselector: selecting attributes. *Vienna: R Foundation for Statistical Computing*.

Schapire, R.E., 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149-171). Springer, New York, NY.

Selvin, E., Steffes, M.W., Zhu, H., Matsushita, K., Wagenknecht, L., Pankow, J., Coresh, J. and Brancati, F.L., 2010. Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *N Engl J Med*, *2010*(362), pp.800-811.

Sharifi, F., Nasab, N.M. and Zadeh, H.J., 2008. Elevated serum ferritin concentrations in prediabetic subjects. *Diabetes and Vascular Disease Research*, *5*(1), pp.15-18.

Shearer, D.M., Thomson, W.M., Broadbent, J.M., McLean, R., Poulton, R. and Mann, J., 2016. High-risk glycated hemoglobin trajectories established by mid-20s: findings from a birth cohort study. *BMJ Open Diabetes Research and Care*, *4*(1), p.e000243.

Strickland, J., 2015. *Predictive analytics using R*. Lulu. com.

Suvitaival, T., Bondia-Pons, I., Yetukuri, L., Pöhö, P., Nolan, J.J., Hyötyläinen, T., Kuusisto, J. and Orešič, M., 2018. Lipidome as a predictive tool in progression to type 2 diabetes in Finnish men. *Metabolism-Clinical and Experimental*, *78*, pp.1-12.

Tabá, G., Herder, C., Rathmann, W., Brunner, E. and Kivimäk, M., 2012. Prediabetes: a high-risk state for developing diabetes. *Lancet*, *379*(9833), pp.2279-2290.

Tao, L.X., Yang, K., Huang, F.F., Liu, X.T., Li, X., Luo, Y.X., Wu, L.J. and Guo, X.H., 2017. Association of Waist Circumference Gain and Incident Prediabetes Defined by Fasting Glucose: A Seven-Year Longitudinal Study in Beijing, China. *International journal of environmental research and public health*, *14*(10), p.1208.

Tripathi, A., 2017. *Practical machine learning cookbook*. Packt Publishing Ltd.

Twig, G., Afek, A., Shamiss, A., Derazne, E., Tzur, D., Gordon, B. and Tirosh, A., 2013. White blood cells count and incidence of type 2 diabetes in young men. *Diabetes care*, *36*(2), pp.276-282.

Urbanowicz, R.J., Meeker, M., LaCava, W., Olson, R.S. and Moore, J.H., 2017. Relief-based feature selection: introduction and review. *arXiv preprint arXiv:1711.08421*.

Villa-Zapata, L., Warholak, T., Slack, M., Malone, D., Murcko, A., Runger, G. and Levengood, M., 2016. Predictive modeling using a nationally representative database to identify patients at risk of developing microalbuminuria. *International urology and nephrology*, *48*(2), pp.249-256.

Warren, B., Pankow, J.S., Matsushita, K., Punjabi, N.M., Daya, N.R., Grams, M., Woodward, M. and Selvin, E., 2017. Comparative prognostic performance of definitions of prediabetes: a prospective cohort analysis of the Atherosclerosis Risk in Communities (ARIC) study. *The Lancet Diabetes & Endocrinology*, *5*(1), pp.34-42.

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one*, *12*(4), p.e0174944.

Wilson, P.W., Meigs, J.B., Sullivan, L., Fox, C.S., Nathan, D.M. and D'Agostino, R.B., 2007. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Archives of internal medicine*, *167*(10), pp.1068-1074.

World Health Organization, 2006. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. *World Hearth Org*.

Zhang, C. and Ma, Y. eds., 2012. *Ensemble machine learning: methods and applications*. Springer Science & Business Media.

Zhang, Q., Bao, X., Meng, G., Liu, L., Wu, H., Du, H., Shi, H., Xia, Y., Guo, X., Liu, X. and Li, C., 2016. The predictive value of mean serum uric acid levels for developing prediabetes. *Diabetes research and clinical practice*, *118*, pp.79-89.

Zhang, Y., Hu, G., Zhang, L., Mayo, R. and Chen, L., 2015. A novel testing model for opportunistic screening of pre-diabetes and diabetes among US adults. *PloS one*, *10*(3), p.e0120382.

Zhao, Y., 2012. *R and data mining: Examples and case studies*. Academic Press.

Zheng, T., Gao, Y., Baskota, A., Chen, T., Ran, X. and Tian, H., 2014. Increased plasma DPP4 activity is predictive of prediabetes and type 2 diabetes onset in Chinese over a four-year period: result from the China National Diabetes and Metabolic Disorders Study. *The Journal of Clinical Endocrinology & Metabolism*, *99*(11), pp.E2330-E2334.

TABLES AND FIGURES

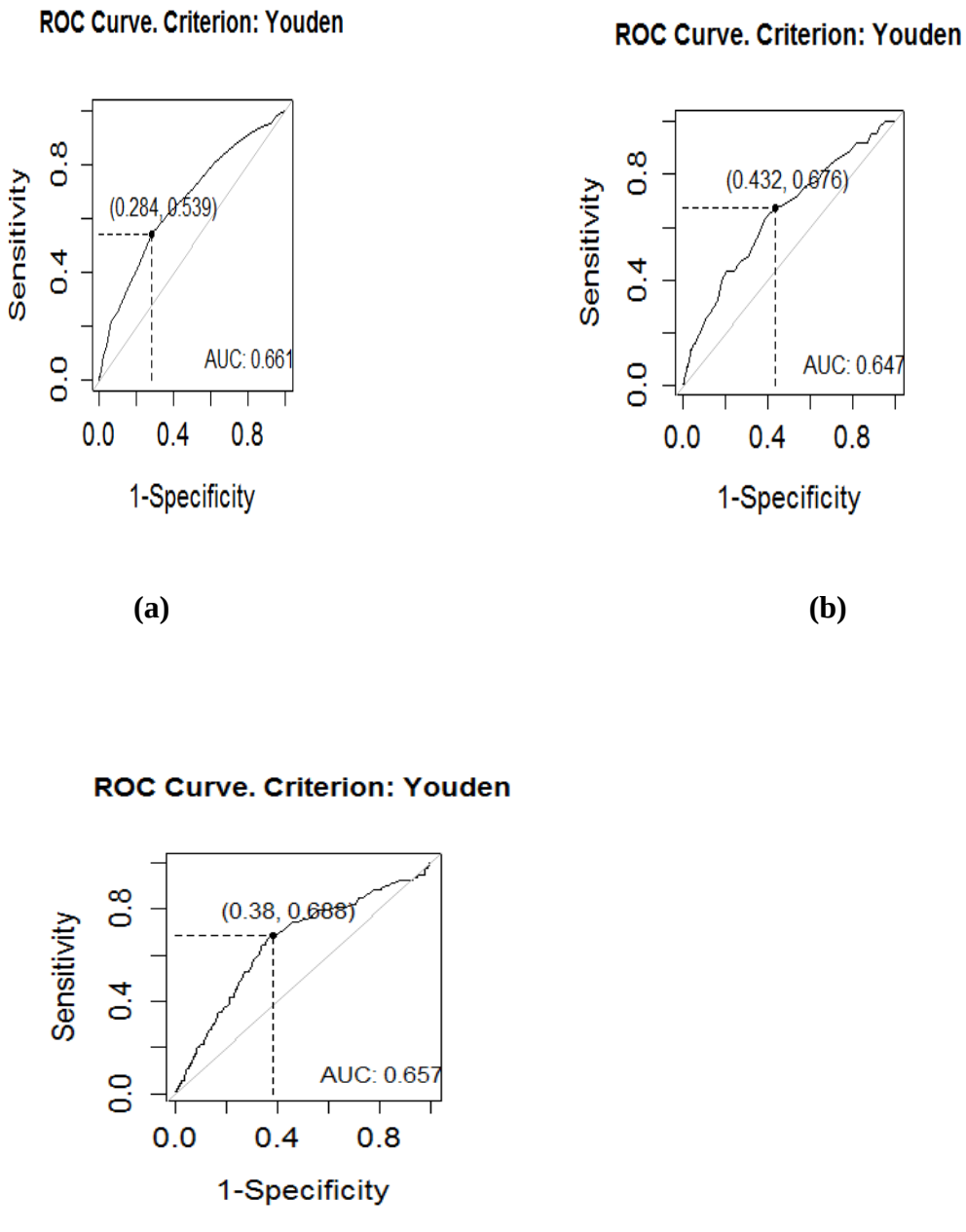


Figure 1: Receiver operating characteristic curves of self-reported status of ever being told prediabetic versus current glycated hemoglobin/Hb1Ac (plot a), fasting plasma glucose/FPG (plot b) and oral glucose tolerance test/OGTT (plot c) among the study sample (N = 6346)

Table 1.1. Distribution of the characteristics of the dataset (N = 6346) containing the 46 extracted general variables of the NHANES 2013-2014 database between prediabetic and non-prediabetic individuals.

Variable	Non-prediabetic (n= 4859)	Prediabetic(n=1487)	p-value*
Categorical variables	**n (%)**	**n (%)**	
Self-perceived DM risk[¥]			
No	3646 (75.04)	1032 (69.40)	**<0.0001**
Yes	1213 (24.96)	455 (30.60)	
Gender[¥]			
Female	2631(54.15)	709 (47.68)	**<0.0001**
Male	2228 (45.85)	778 (52.32)	
Race[¥]			
(Non-Hispanic) White	1905 (39.21)	602 (40.48)	NS
Other[a]	2954 (60.79)	885 (59.52)	
Citizenship[¥]			
Yes	4290 (88.29)	1288 (86.62)	NS
No	569 (11.71)	199 (13.38)	
Marital status[¥]			
Married/ Living with partner	3564 (73.35)	1063 (71.49)	NS
Other[b]	1295 (26.65)	424 (28.51)	
Alcohol use[c, ¥]			
No	1465 (30.15)	452 (30.40)	NS
Yes	3394 (69.85)	1035 (69.60)	
Past any tobacco use[¥]			
No	3757 (77.32)	1164 (78.28)	NS
Yes	1102 (22.68)	323 (21.72)	
Diagnosed hypertension[¥]			
No	3573 (73.53)	890 (59.85)	**<0.0001**
Yes	1286 (26.47)	597 (40.15)	
Hepatitis B[¥]			
Yes	41 (0.84)	16 (1.08)	NS
No	4818 (99.16)	1471 (98.92)	
Hepatitis C[¥]			
Yes	48 (0.99)	18 (1.21)	NS
No	4811 (99.01)	1469 (98.79)	
Diagnosed jaundice[¥]			
No	4759 (97.94)	1458 (98.05)	NS
Yes	100 (2.06)	29 (1.95)	
Familial diabetes[¥]			
No	3025 (62.26)	900 (60.52)	NS
Yes	1834 (37.74)	587 (39.48)	
Hepatitis B surface antibody			
Negative	3418 (70.34)	1153 (77.54)	**<0.0001**
Positive	1441 (29.66)	334 (22.46)	
Hepatitis E IgG			
Negative	4661 (95.93)	1400 (94.15)	**0.0038**
Positive	198 (4.07)	87 (5.85)	
Education[¥,d]			
<9th grade	733 (15.08)	207 (13.92)	NS
9-11 grade	985 (20.27)	256 (17.22)	**0.0093**
High school	944 (19.43)	306 (20.58)	NS
College/AA degree/above	2197 (45.22)	718 (48.28)	**0.0377**
Vigorous activity[t]			
No	2762 (56.84)	948 (63.75)	**<0.0001**

Yes	2097 (43.16)	539 (36.25)	
Moderate activity[ŧ]			
No	1383 (28.46)	470 (31.61)	**0.0196**
Yes	3476 (71.54)	1017 (68.39)	
Gestational DM[¥]			
No	4731 (97.37)	1443 (97.04)	NS
Yes	128 (2.63)	44 (2.96)	
Overweight baby at birth (> 9lb) [¥]			
No	4635 (95.39)	1384 (93.07)	**0.0004**
Yes	224 (4.61)	103 (6.93)	
Hysterectomy[¥]			
No	4514 (92.90)	1307 (87.90)	**<0.0001**
Yes	345 (7.10)	180 (12.10)	
Bilateral ovariectomy[¥]			
No	4669 (96.09)	1396 (93.88)	**0.0003**
Yes	190 (3.91)	91 (6.12)	
Female hormones intake[¥]			
No	4537 (93.37)	1327 (89.24)	**<0.0001**
Yes	322 (6.63)	160 (10.76)	
Numeric variables			
Variable	**Mean (SD)**	**Mean; SD**	**p-value***
Age (years) [¥]	38.18 (20.00)	48.87 (19.74)	**<0.0001**
Income-poverty ratio	2.43 (1.65)	2.46 (1.63)	NS
Food security[¥,d]	3.50 (0.91)	3.37 (0.99)	**<0.0001**
Duration of watching TV (hours) [¥]	2.29 (1.63)	2.50 (1.62)	**<0.0001**
Body mass index (kg/m^2)	27.21 (6.87)	29.43 (7.54)	**<0.0001**
Waist circumference (cm)	92.94 (17.10)	99.98 (17.28)	**<0.0001**
White cell count(×10^9/L)	7.14 (2.20)	7.25 (2.29)	NS
Monocyte count(×10^9/L)	0.579 (0.20)	0.577 (0.19)	NS
Red cell count (million cells/uL)	4.65 (0.48)	4.75 (0.51)	**<0.0001**
Hemoglobin(g/dL)	13.90 (1.47)	14.19 (1.54)	**<0.0001**
Alanine aminotransferase (U/L)	24.49 (13.51)	26.10 (26.74)	**0.0019**
Aspartate aminotransferase (U/L)	22.95 (17.67)	25.83 (19.66)	**<0.0001**
Serum calcium (mg/ dL)	9.46 (0.35)	9.50 (0.37)	**0.0002**
Serum globulin (g/dL)	2.80 (0.43)	2.85 (0.43)	**<0.0001**
Gamma glutamyl transferase (U/L)	23.40 (33.29)	27.41 (34.04)	**<0.0001**
Serum iron (ug/dL)	83.78 (37.73)	85.40 (34.41)	NS
Serum potassium (mmol/L)	4.07 (0.36)	4.00 (0.34)	**<0.0001**
Osmolality (mmol/kg)	278.60 (4.63)	279.64 (4.88)	**<0.0001**
Serum phosphorus (mg/dL)	3.98 (0.65)	3.85 (0.65)	**<0.0001**
Triglycerides (mg/dL)	133.89 (97.85)	135.50 (96.65)	NS
Serum uric acid (mg/dL)	5.22 (1.35)	5.60 (1.42)	**<0.0001**
Mean SBP (mmHg)	118.00 (16.89)	123.66 (17.82)	**<0.0001**
Mean DBP (mmHg)	66.55 (13.03)	67.81 (13.26)	**0.0012**
Hematocrit	41.11 (3.96)	42.06 (4.20)	**<0.0001**

*** Chi-squared test for 2 proportions and 2-samples t-test were used for univariate analyses of categorical and continuous variables, respectively. Level of significance p = 0.05, NS-not significant, a = Mexican American, other Hispanic, non-Hispanic Black, non-Hispanic Asian & other races including multi-racial, b = widowed, divorced, separated or never-married, c = defined as use of at least 12 drinks of any alcoholic beverage in any 1 year, ¥ = self-reported data, ŧ = composite variables derived using NHANES questionnaire, d = modelled as continuous variables, DM = diabetes mellitus, SBP = systolic blood pressure, DBP = diastolic blood pressure, SD = standard deviation, IgG = immunoglobulin G, AA degree = Associate of Arts degree, equivalent to the first two years of a bachelor's degree**

Table 1.2. Distribution of the characteristics of the dataset containing dental variables of the NHANES 2013-2014 between prediabetic and non-prediabetic individuals. Chi-squared test for two proportions and two-samples t-test were used for univariate distributional analyses of categorical and continuous variables, respectively.

Variable	Non-prediabetic (n = 2234)	Prediabetic (n = 933)	p-value*
Numeric variables			
Variable	**Mean (SD)**	**Mean; SD**	
Age (years) ¥	49.92 (14.10)	53.71 (14.22)	**<0.0001**
Income-poverty ratio	2.74 (1.68)	2.59 (1.66)	**0.0216**
Time since last dental visit[¥,e]	2.51 (1.83)	2.63 (1.89)	NS
Self-rated oral health[¥,f]	2.96 (1.11)	3.04 (1.14)	NS
Dental floss/device use frequency[¥,g]	3.46 (2.91)	3.46 (2.97)	NS
Mouthwash use frequency[¥,h]	2.99 (3.11)	3.07 (3.13)	NS
Number of teeth	24.56 (6.23)	23.81 (6.80)	**0.0027**
Categorical variables	**n (%)**	**n (%)**	
Periodontitis[a]			
No	1413 (63.25)	555(59.49)	**0.0465**
Yes	821 (36.75)	378 (40.51)	
Gender¥			
Male	1020 (45.66)	479 (51.34)	**0.0035**
Female	1214 (54.34)	454 (48.66)	
Race¥			
(Non-Hispanic) White	1006 (45.03)	399 (42.77)	NS
Other[b]	1228 (54.97)	534 (57.23)	
Citizen of United States¥			
Yes	1926 (86.21)	802 (85.96)	NS
No	308 (13.79)	131 (14.04)	
Marital status¥			
Married/ Living with partner	1595 (71.40)	665 (71.28)	NS
Other[c]	639 (28.60)	268 (28.72)	
Ever had periodontal treatment¥			
Yes	540 (24.17)	218 (23.37)	NS
No	1694 (75.83)	715 (76.63)	
Self-reported tooth mobility¥			
Yes	303 (13.56)	156 (16.72)	**0.0214**
No	1931 (86.44)	777 (83.28)	
Education[d, ¥]			
<9th grade	141 (6.31)	83 (8.90)	**0.0097**
9-11 grade	243 (10.88)	128 (13.72)	**0.0234**
High school	473 (21.17)	202 (21.65)	NS
College/AA degree/ college graduate or above	1377 (61.64)	520 (55.73)	**0.0020**

NS-not significant, *Level of significance was set at p = 0.05, a = see no:8 of the table 5 in the appendix for details, b = Mexican American, other Hispanic, non-Hispanic Black, non-Hispanic Asian & other races including multi-racial, c = widowed, divorced, separated and never-married, d = modelled as a continuous variable, SD = standard deviation, AA degree = Associate of Arts degree, equivalent to the first two years of a bachelor's degree, ¥ = self-reported data, e- entered as a continuous variable scaled 1-7; see appendix: table 5 (no:5), f- entered as a continuous variable scaled 1-5; see appendix: table 5 (no:3), g- number of days during the past week dental floss was used, h- number of days during the past week a mouthwash/oral rinse was used

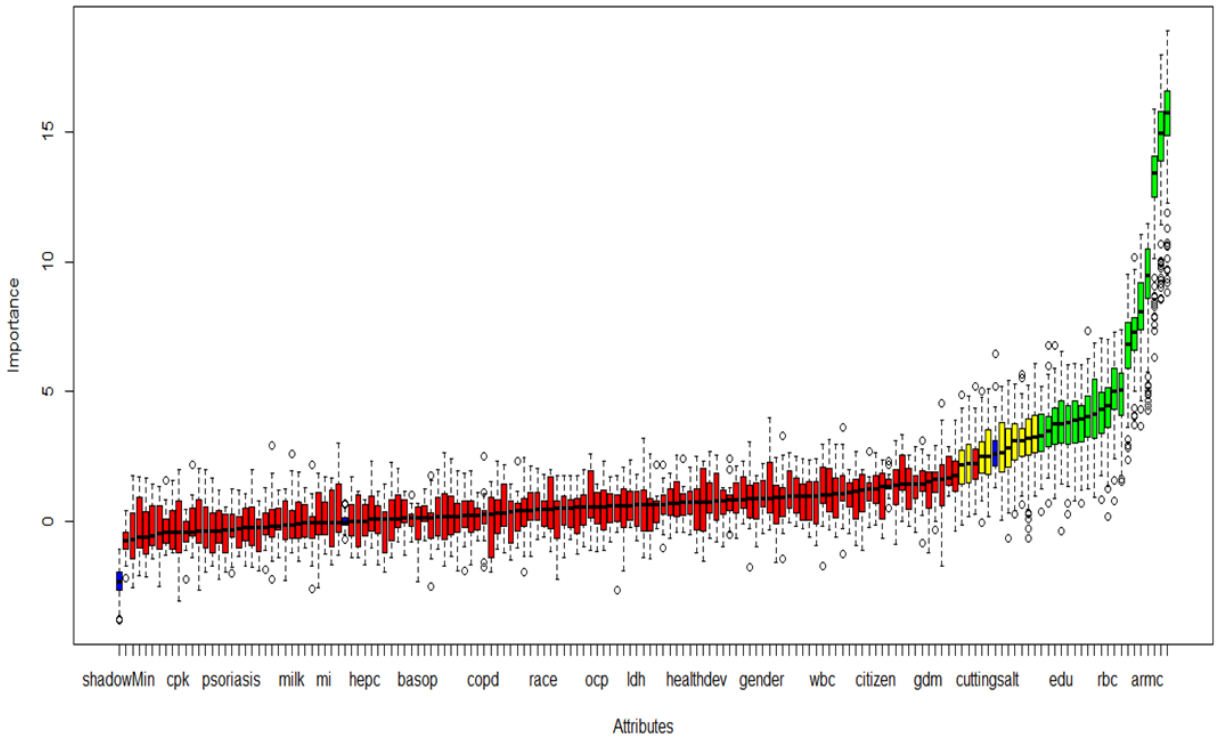
Table 2: A summary of the feature selection algorithms employed on the train dataset (n = 3134) containing the 156 general variables and the attributes selected by each algorithm

Wrapper algorithms		
package	**Feature selection algorithm**	**Extracted variables**
"Boruta" (Kursa & Rudnicki, 2010)	An all-relevant feature selection algorithm using a random forest classifier. The algorithm consists of several steps summarized below (Kursa & Rudnicki, 2010): *First, randomness is added to the dataset by creating shuffled copies of all features which are also called shadow features. *Then, a random forest classifier is run on the extended data set and a feature importance measure (the default is mean decrease accuracy) is applied to evaluate the importance of features. *At every iteration, the algorithm checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed unimportant. *Finally, the algorithm stops either when all features get confirmed or rejected or it reaches a specified limit of random forest runs.	confirmed (20): **age, marital status**, advised self-measured BP[a], advised to lose weight, advised to reduce salt intake, advised to reduce fat intake, **BMI[b]**, arm circumference, **waist circumference**, sagittal abdominal diameter, **red cell count, hemoglobin, osmolality, triglyceride level, education, bilateral ovariectomy, female hormones intake, mean SBP[c], mean DBP[d], hematocrit** tentative (10): reducing salt intake, advised to increase exercise, diagnosed hypercholesterolemia, **GGT[e], hepatitis E IgG[f]**, served in armed forces, **diagnosed hypertension, serum potassium level, serum uric acid, hysterectomy**
Filter algorithms		
"FSelector" (Romanski & Kotthoff, 2009)	The package contains both filter and wrapper functions but only the filter functions were used in the present study, namely, gain ratio, symmetrical uncertainty, random forest, and relief. For using entropy-based methods, continuous features were discretized.	
	1)Gain ratio: An entropy-based filter based on information gain criterion derived from a decision-tree classifier modified to reduce bias on highly branching features with many values. Bias reduction is achieved through normalizing information gain by the intrinsic information of a split (Harris, 2002).	Top 30: **age, waist circumference**, sagittal abdominal diameter, **monocyte count, mean SBP[c]**, diagnosed hypercholesterolemia, **BMI[b], diagnosed hypertension, uric acid, GGT[e], serum phosphorus**, arm circumference, **vigorous activity**, **familial diabetes**, **marital status**, advised to reduce salt intake, advised to lose weight, advised self-measured BP, **serum potassium** , self-measured BP, **hepatitis B**, **hepatitis C, race, ALT[g], overweight baby at birth, gender, hepatitis B surface antibody**, advised to reduce fat intake, **bilateral ovariectomy, self-perceived DM risk[h]**
	2)Symmetrical uncertainty: An entropy-based filter based on information gain criterion but modified to reduce bias on highly branching features with many values. Bias reduction is achieved through normalizing information gain by the corresponding entropy of features (Press et al., 1996).	Top 30: **age, waist circumference**, sagittal abdominal diameter, **mean SBP[c]**, diagnosed hypercholesterolemia, **BMI[b], gender, GGT[e], race, serum uric acid, phosphorus**, arm circumference, **hepatitis E IgG[f]**, advised to increase exercise, **hepatitis B, serum potassium, food security**, advised to lose weight, advised

		to reduce salt intake, **ALT[g]**, self-measured BP, **hepatitis B surface antibody**, advised to reduce fat intake, **hepatitis C**, advised self-measured BP, **self-perceived DM risk[h]**, self-rated general health, self-rated health trend, **female hormone intake, hysterectomy**
	3)Random forest: The algorithm finds weights of attributes using random forest algorithm. Random forest is an ensemble learner with additional randomness incorporated to independently-constructed and bootstrap aggregated classification trees (Liaw & Wiener, 2002).	Top 30: **age, waist circumference**, sagittal abdominal diameter, **BMI[b], mean SBP[c]**, arm circumference, **mean DBP[d], income-poverty ratio, hematocrit, osmolality, triglycerides level, bilateral ovariectomy, RBC count[i], female hormones intake**, poverty index, advised self-measured BP[a], advised to reduce salt intake, **WBC[j] count**, diagnosed hypercholesterolemia, advised to reduce fat intake, **marital status, serum uric acid, hemoglobin, GGT[e], monocyte count, serum calcium, hepatitis E IgG[f]**, annual household income, **serum phosphorus**, family income
	4)Relief: The algorithm finds weights of continuous and discrete attributes basing on a distance between instances (Urbanowicz et al, 2017).	Top 30: **education,** managing weight, **past any tobacco use**, duration of computer use, health insurance availability, **hepatitis C**, poverty category, reducing fat intake, diagnosed asthma, **vigorous activity**, self-rated general health, **overweight baby at birth**, **citizenship**, reducing salt intake, **alcohol use**, **female hormone intake**, **moderate activity**, **hysterectomy**, nightly urinating frequency, **duration of watching TV**, annual family income, annual household income, **bilateral ovariectomy**, **gestational DM[h]**, advised to reduce fat intake, monthly family income, **diagnosed jaundice**, duration of sleep, **familial diabetes. Hepatitis E IgG[f]**
Embedded algorithms		
"glmnet" (Friedman et al, 2010)	Lasso (Least Absolute Shrinkage and Selection Operator) regularization: This puts a constraint on the sum of the absolute values of the model parameters. The sum should be less than a fixed value (upper bound). A shrinking (regularization) process is used where it penalizes the coefficients of the regression variables shrinking some of them to zero. The variables that have a non-zero coefficient after the shrinking process are selected. The process minimizes the prediction error (Fonti & Belitser, 2017).	15 features: **self-perceived DM risk[h], age, citizenship, diagnosed hypertension**, diagnosed hypercholesterolemia, self-rated health trend, advised to increase exercise, **waist circumference**, sagittal abdominal diameter, **RBC[i] count, hepatitis E IgG[f]**, **serum iron**, **serum calcium, serum globulin, serum potassium**
"caret" (Kuhn et al, 2017)	Recursive feature elimination: A resampling based recursive feature elimination method is applied by default by the caret package. A random forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes	Top 30: **age, waist circumference**, sagittal abdominal diameter, **duration of watching TV, mean SBP[c], hematocrit, WBC[j] count, GGT[e], gestational DM[h], mean DBP[d]**, arm circumference, **hepatitis**

(Kuhn et al, 2017).	**E IgG[f], income-poverty ratio, food security**, diagnosed hypercholesterolemia, **RBC[i] count, marital status, osmolality, diagnosed jaundice**, pulse character, advised self-measured BP[a], annual household income, **serum uric acid, overweight baby at birth, serum iron**, advised to reduce fat intake, **BMI[b], AMT[k], hysterectomy**, advised to lose weight

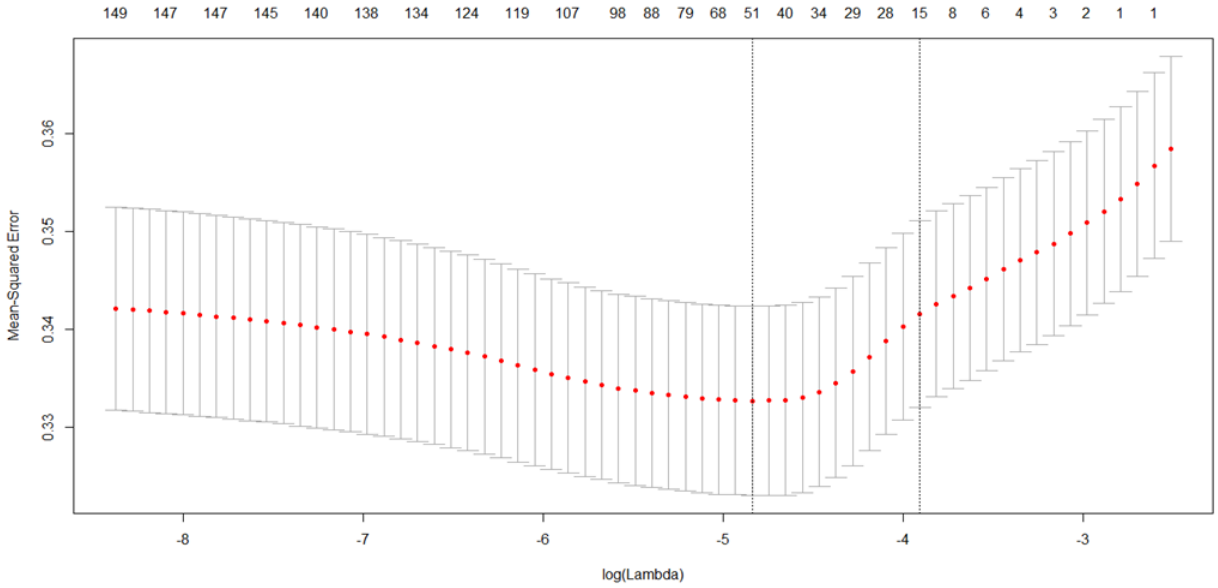
a-BP=blood pressure, b-BMI=body mass index, c-SBP=systolic blood pressure, d-DBP= diastolic blood pressure, e-GGT=gamma glutamyl transferase, f-IgG=immunoglobulin G, g-ALT=alanine amino transferase, h-DM=diabetes mellitus, i-RBC=red blood cells, j-WBC=white blood cells, k-AMT=aspartate amino transferase



(shadowMin=Minimum shadow score, cpk=creatine phosphokinase, psoriasis=diagnosed psoriasis, milk=milk consumption, mi=diagnosed heart attack, hepc=hepatitis C, basop=basophil count, copd=diagnosed chronic obstructive pulmonary disease, ocp=oral contraceptive use, ldh=lactate dehydrogenase, healthdev=self-rated health trend, wbc=white cell count, citizen=citizenship status, gdm=gestational diabetes, cuttingsalt=reducing salt intake, edu=education, rbc=red cell count, armc=arm circumference).

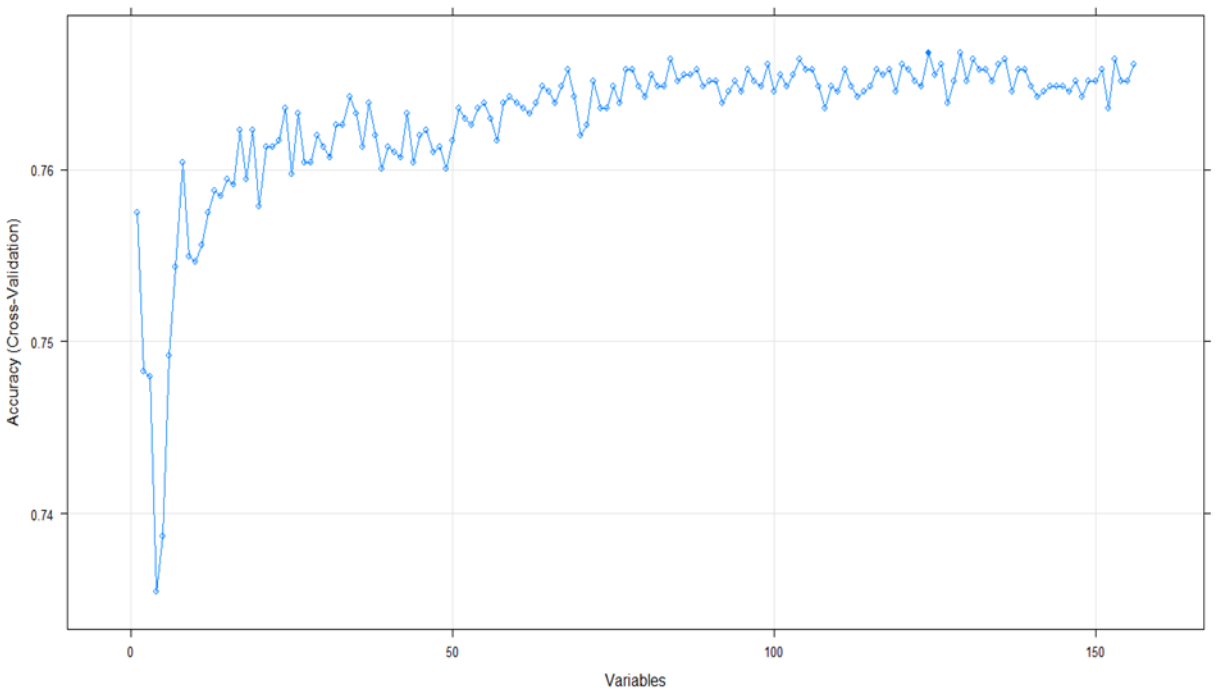
Default functions of the "Boruta" R package were used; feature importance measure = mean decrease accuracy, maximal number of random forest runs = 100. Red, yellow, green, and blue boxplots represent Z scores of rejected, tentative, confirmed and shadow attributes respectively. Shadow (minimum, mean, and maximum) features are reference points for deciding which attributes are truly important and these values are generated by the algorithm via shuffling values of the original attributes (Kursa & Rudnicki, 2010). The 20 confirmed and the 10 tentative features selected by the "Boruta" algorithm are given in the table 2.

Figure 2.1. Feature selection using Boruta algorithm: Variable importance plot



The lambda value that minimizes the cross validated mean squared error determines the sparse model containing the selected features. Default functions of the "glmnet" R statistical package were used. Fifteen features selected by the algorithm are given in the table 2.

Figure 2.2. Feature selection using Lasso regularization



A random forest classifier with two-fold cross-validation was specified with other default functions of the "caret" package in R to extract features via recursive feature elimination. The 30 most important features selected by the recursive feature elimination algorithm are given in the table 2.

Figure 2.3. Feature selection using recursive feature elimination

Objective1

Table 3.1: Determinants of prediabetes elucidated by predictive models with an AUC > 70% built using logistic regression algorithm

Logistic regression models and elucidated determinants					
GLM original[a] ($AUC_{internal}$ = 70.76%) ($AUC_{external}$=69.56%)		**GLM undersampled[b] ($AUC_{internal}$ = 70.30%) ($AUC_{external}$ = 69.01%)**		**GLM oversampled[c] ($AUC_{internal}$ = 70.83%) ($AUC_{external}$ = 69.62%)**	
Predictor	**OR (95% CI)**	**Predictor**	**OR (95% CI)**	**Predictor**	**OR (95% CI)**
Socio-economic		**Socio-economic**		**Socio-economic**	
Age	1.02 (1.01-1.03)	Age	1.02 (1.01-1.03)	Age	1.02 (1.01-1.03)
Citizenship (reference = yes)	1.38 (1.04-1.81)	**Clinical**		Citizenship (reference = yes)	1.23 (1.01-1.50)
Marital status (reference = unmarried)	0.95 (0.90-1.00)	Hepatitis C (reference = no)	1.20 (1.02-1.43)	Marital status (reference = unmarried)	0.95 (0.91-0.98)
Income-poverty ratio	0.94 (0.88-1.00)	**Biochemical**		Income-poverty ratio	0.92 (0.88-0.96)
Food security	0.85 (0.76-0.94)	Monocyte count	0.44 (0.20-0.94)	Food security	0.82 (0.77-0.89)
Clinical		Serum potassium	0.60 (0.43-0.83)	**Clinical**	
Diagnosed HT (reference = no)	1.26 (1.02-1.55)	Uric acid	1.14 (1.03-1.26)	Diagnosed HT (reference = no)	1.18 (1.02-1.37)
Mean SBP	1.01 (1.00-1.02)			Vigorous exercise (reference = no)	0.47 (0.28-0.76)
Biochemical				Hysterectomy (reference = no)	1.42 (1.04-1.93)
Monocyte count	0.45 (0.24-0.82)			**Biochemical**	
Red cell count	1.51 (1.14-2.01)			GGT[e]	1.10 (1.00-1.20)
Serum calcium	1.39 (1.06-1.82)			Monocyte count	0.43 (0.28-0.65)
ALT[d]	1.33 (1.07-1.65)			Red cell count	1.32 (1.07-1.62)
Serum potassium	0.58 (0.45-0.75)			Serum calcium	1.39 (1.15-1.67)
Triglycerides	1.01 (1.00-1.02)			ALT[d]	1.46 (1.24-1.71)
				Serum potassium	0.58 (0.49-0.70)
				Osmolality	1.02 (1.00-1.03)
				Uric acid	1.11 (1.04-1.20)
				Triglycerides	1.01 (1.00-1.02)
				Hematocrit	1.09 (1.03-1.14)

a=logistic regression model on original, un-resampled data, b= logistic regression model on the train data re-structured by majority class under-sampling, c= logistic regression model on the train data re-structured by minority class oversampling, AUC=area under receiver operating characteristic curve, $AUC_{internal}$=AUC on the internal validation data, $AUC_{external}$=AUC on the external validation data, OR=odds ratio, CI=confidence interval, HT=hypertension, SBP=systolic blood pressure, d=serum alanine amino-transferase, e=serum gamma glutamyl transferase

Table 3.2: Determinants of prediabetes elucidated by predictive models with an AUC > 70% built using non-linear and ensemble machine learning algorithms. Importance values of the 20 most influential socio-economic, clinical, and biochemical predictors of each model are given in descending order.

Optimal non-linear/ensemble machine learning models and elucidated determinants							
RF oversampled[a] ($AUC_{internal}$ =71.59%) ($AUC_{external}$ =70.01%)		**RF SMOTE[b] ($AUC_{internal}$ = 70.66%) ($AUC_{external}$ =69.23%)**		**ANN original[c] ($AUC_{internal}$ =70.21%) ($AUC_{external}$ =68.95%)**		**XGB original[d] ($AUC_{internal}$ = 70.55%) ($AUC_{external}$ = 69.45%)**	
Predictor	**importance[e]**	**Predictor**	**importance[f]**	**Predictor**	**importance[g]**	**Predictor**	**importance[h]**
Socio-economic		**Socio-economic**		**Socio-economic**		**Socio-economic**	
Age	115.16	Age	113.93	Age	0.6457	Age	100.00
Income-poverty ratio	75.71	Income-poverty ratio	74.85	Marital status	0.5848	Food security	5.12
Clinical		**Clinical**		Food security	0.5363	Citizenship	2.36
Waist circumference	103.60	Waist circumference	101.23	**Clinical**		**Clinical**	
Body mass index	94.50	Body mass index	90.78	Mean SBP	0.5961	Waist circumference	35.12
Mean SBP	90.44	Mean SBP	88.20	Body mass index	0.5888	Hepatitis C	8.76
Diagnosed HT	81.26	Hepatitis C	81.25	Diagnosed HT	0.5668	Body mass index	4.78
Hepatitis C	79.98	Hepatitis B	74.55	Hepatitis C	0.5554	Mean SBP	4.27
Hepatitis B	78.35	Vigorous exercise	73.52	Hepatitis B	0.5548	Hepatitis B	2.82
Vigorous exercise	67.07	Diagnosed HT	72.95	Vigorous exercise	0.5387	Diagnosed HT	2.23
Biochemical		**Biochemical**		Waist circumference	0.5381	**Biochemical**	
Red cell count	87.82	GGT[i]	88.69	Hysterectomy	0.5376	Serum potassium	15.13
Triglycerides	86.61	Serum potassium	86.86	**Biochemical**		Red cell count	14.48
Serum potassium	85.63	Serum calcium	82.67	GGT[i]	0.5894	Triglycerides	13.97
GGT[i]	84.45	Uric acid	80.70	Uric acid	0.5812	Hematocrit	11.77
Serum calcium	83.72	Osmolality	79.46	Serum potassium	0.5765	GGT[i]	10.49
Uric acid	82.39	Triglycerides	79.46	ALT[j]	0.5731	Osmolality	8.18
White cell count	80.59	Monocyte count	78.14	AMT[k]	0.5640	Uric acid	6.91
ALT[j]	75.74	ALT[j]	77.35	Hematocrit	0.5636	White cell count	5.57
Osmolality	74.55	Red cell count	77.18	Osmolality	0.5630	ALT[j]	4.54
AMT[k]	70.65	Hematocrit	77.17	Red cell count	0.5517	AMT[k]	4.37
Hematocrit	70.55	White cell count	74.72	Triglycerides	0.5357	Serum calcium	2.25

a=random forest model on train data restructured by minority class oversampling, b= random forest model on train data restructured by synthetic minority oversampling algorithm, c=artificial neural network model on original, un-resampled train data, d=gradient boosting model on original, un-resampled train data, e,f= by default, mean decrease in prediction accuracy after a variable is permuted, g= default method is based on Gevrey et al (2003), which uses combinations of the absolute values of the weights, h=same approach as a single tree (i.e. reduction in the loss function attributed to each variable at each split is tabulated, summed over each node, and totaled.) but sums the importance estimates over each boosting iteration, AUC=area under receiver operating characteristic curve, $AUC_{internal}$=AUC on the internal validation data, $AUC_{external}$=AUC on the external validation data, HT=hypertension, SBP=systolic blood pressure, i=gamma glutamyl transferase, j=alanine amino-transferase, k=aspartate aminotransferase

Objective 2

Table 4: Determinants of prediabetes elucidated by optimal predictive models built using dental variables and machine learning algorithms

Optimal model using each algorithm and elucidated determinants				
GLM ROSE[a]	**ANN ROSE[b]**	**RF ROSE[c]**	**XGB under-sampled[d]**	**Bagged CART original[e]**
$AUC_{internal}$ =59.43%	**$AUC_{internal}$=58.21%**	**$AUC_{internal}$=58.91%**	**$AUC_{internal}$= 58.38%**	**$AUC_{internal}$ = 54.53%**
$AUC_{external}$=57.88%	**$AUC_{external}$=56.93%**	**$AUC_{external}$=57.29%**	**$AUC_{external}$= 57.14%**	**$AUC_{external}$= 53.38%**
Variable; OR (95% CI)	**Variable; importance[f]**	**Variable; importance[g]**	**Variable; Importance[h]**	**Variable; Importance[i]**
Age 1.02 (1.01-1.03)	Age; 0.5687	Age; 65.74	Age; 100.00	Age; 100.00
Periodontitis* (reference = no); 1.43 (1.12-1.82)	Race; 0.5380	Periodontitis*; 57.79	Income-poverty ratio; 26.73	Income-poverty ratio; 91.50
Mobile teeth (reference = no); 1.59 (1.14-2.17)	Periodontitis*; 0.5354	Education**; 57.56	Periodontitis*; 19.78	Periodontitis*; 68.98
Gender (reference = female); 1.47 (1.19-1.82)	Education**; 0.5270	Gender; 55.51	Race; 14.53	Race; 42.41
Education**; 0.87; (0.77-0.99)	Gender; 0.5256	Race; 52.77	Gender; 9.66	Education**; 41.94

a=logistic regression on train data restructured by random oversampling(ROSE) algorithm, b=artificial neural network model on train data restructured by random oversampling(ROSE) algorithm, c=random forest model on train data restructured by random oversampling(ROSE) algorithm, d=gradient boosting model on train data re-structured by majority class under-sampling, e=Bagged CART model on original, un-resampled train data, f= default method is based on Gevrey et al (2003), which uses combinations of the absolute values of the weights, g= by default, mean decrease in prediction accuracy after a variable is permuted, h=same approach as a single tree (i.e. reduction in the loss function attributed to each variable at each split is tabulated, summed over each node, and totaled.) but sums the importance estimates over each boosting iteration, i=The same methodology as a single tree (i.e. reduction in the loss function attributed to each variable at each split is tabulated and the sum is returned) is applied to all bootstrapped trees and the total importance is returned, OR=odds ratio, CI=confidence interval, AUC=area under receiver operating characteristic curve, $AUC_{internal}$=AUC on the internal validation data, $AUC_{external}$=AUC on the external validation data, *classified as per Eke et al (2012) and dichotomized; see no:8 of the table 5 in the appendix for details, **modelled as a continuous variable

Objective 3

Table 5: Comparison of the performance of the Centers for Disease Control and prevention (CDC) prediabetes screening tool upon the NHANES database with that of optimal predictive models

Benchmarking with the optimal model containing general predictors*				
Criterion	**CDC prediabetes screening tool**		**Optimal machine learning model***	
	Performance upon the internal validation dataset[a] (N = 3172)	**Performance upon the external validation dataset[b] (N = 3000)**	**Performance upon the internal validation dataset[a] (N = 3172)**	**Performance upon the external validation dataset[b] (N = 3000)**
AUC	64.40%	62.80%	71.59%	70.01%
Sensitivity	63.53%	62.50%	71.87%	72.33%
Specificity	59.78%	59.55%	61.67%	59.22%
Accuracy	61.07%	60.24%	63.43%	75.68%
PPV	32.57%	32.11%	34.72%	36.88%
NPV	84.27%	83.84%	86.57%	85.40%
Kappa	0.1766	0.1660	0.2175	0.1543
Benchmarking with the optimal model containing dental predictors**				
Criterion	**CDC prediabetes screening tool**		**Optimal machine learning model****	
	Performance upon the internal validation dataset[a] (N = 1583)	**Performance upon the external validation dataset[b] (N = 1500)**	**Performance upon the internal validation dataset[a] (N = 1583)**	**Performance upon the external validation dataset[b] (N = 1500)**
AUC	59.10%	58.80%	59.43%	57.88%
Sensitivity	71.28%	70.24%	71.24%	68.93%
Specificity	42.84%	42.61%	42.52%	44.80%
Accuracy	51.56%	51.14%	59.13%	58.95%
PPV	34.24%	33.85%	34.09%	34.58%
NPV	78.12%	77.40%	78.00%	79.51%
Kappa	0.1080	0.0980	0.1448	0.1549

* This was a random forest model on train data restructured by minority class oversampling, ** This was a logistic regression model on train data restructured by random oversampling(ROSE) algorithm, a=from NHANES 2013-2014, b=from NHANES 2011-2012, AUC= area under the receiver operating characteristic curve, PPV=positive predictive value, NPV= negative predictive value

Table 6: Derivation of the variables of the Centers for Disease Control and prevention (CDC) prediabetes screening tool using the National Health and Nutrition Examination Survey (NHANES) database and corresponding parameter scores applied

Variable	How information was retrieved/derived from the NHANES database	Score
Age < 65 years & physically inactive	Age was categorized with the cut-points of 45 and 65. See below for "physically inactive"	5
Age 45 – 64 years		5
Age ≥ 65 years		9
Parent DM Sibling DM	NHANES questionnaire collects information on familial DM but not separately for parent's and sibling's DM so the 2 questions were combined and assigned the score of 2	2
BMI ≥ 27 kg/m^2	The CDC score provides a table of weight and height where the classification corresponds to BMI cut-point of 27 kg/m^2 (2 groups)	5
Physically inactive	Derived a binary variable by checking if any of the following activities were done in 5 or more days of a typical week: vigorous or moderate work, recreational work, walk or bicycle	See "age" above
Baby > 9 pounds birth weight (women only)	Available in the questionnaire	1

DM= diabetes mellitus, BMI = body mass index, NHANES = National Health and Nutrition Examination Survey

APPENDIX

Table 1: A summary review of predictive modelling studies for prediabetes, diabetes related conditions and phenomena

Study	Condition /phenomenon	Predictors	Modelling technique	Key facts and findings
Heikes et al, 2008	Diabetes and prediabetes	18 established predictors requiring no lab tests	Logistic regression and CART	NHANES III (N = 7092) was used as the database. External validation was done against NHANES 1999–2004 data. The AUC for prediabetes or undiagnosed diabetes was found to be 0.75. The study proposed the concept of "diabetes risk calculator"; a noninvasive screening tool to detect both pre-diabetes and undiagnosed diabetes in the US population.
Ogunyemi et al, 2015	Diabetic retinopathy	11 clinical predictors from a pool of 24 and 10 public health predictors from a pool of 33 were selected using Lasso regularization	Ensemble classifiers with & without adjustments for class imbalance	Two databases, namely, clinical data from urban safety net clinics (N = 513) & public health data from NHANES 2005-2008 (N = 2874) were used. Only clinical data predicted disease; best model had an AUC of 0.72. Classifiers on NHANES 2005-2008 data were not predictive of diabetic retinopathy.
Choi et al, 2014	Prediabetes	9 established risk factors were used based on literature review	ANN and SVM	Data from the Korean National Health and Nutrition Examination Survey (KNHANES) 2010 (n = 4685) were used for training and internal validation whilst data from KNHANES 2011 (n = 4566) were used for external validation of models. In addition, model performance was compared with that of a screening score tool used among the Koreans. Two optimal models were reported; an ANN model with an AUC of 0.768 and a SVM model with an AUC of 0.761 on the internal validation data. The AUC of the screening score on the internal validation data was 0.734.
Lee et al, 2012	Diabetes	6 established predictors were used for the screening model	Logistic regression	*Data from KNHANES 2001 & 2005 (N = 9602) were used for building the predictive model and the external validation was performed against KNHANES 2007-2008 data. The developed screening tool for undiagnosed diabetes was compared with other similar tools. The AUC of the model was 0.73.
Baan et al, 1999	Diabetes	15 established predictors in total in 3 different models	Logistic regression	A sample of participants from the Rotterdam Study (n = 1016) was used for building the models. The sample consisted of people aged 55–75 years. The external validation was performed against another Dutch population-based study, the Hoorn Study (n = 2364). The AUC of the best model was 0.74.
Anderson et al, 2016	Progression to prediabetes and diabetes	442 variables spanning demographics, lab tests, clinical examinations, ICD-9 diagnosis codes and prescriptions were used	A novel ensemble analytic platform; Reverse Engineering and Forward Simulation (REFS) Bayesian Analytics Platform	Electronic health records data of 24 331 retrospectively followed patients were used for building the prediction models. The optimal model had an AUC of 0.76. Models of progression to prediabetes elucidated novel predictors whilst models of progression to diabetes elucidated established predictors.

Dall et al, 2014	Diabetes and prediabetes	16 variables including 2 potentially new risk factors; history/previous diagnosis of asthma and arthritis were used	Multinomial and binary logistic regression	Predictive modelling was performed on NHANES 2003–2010 (n = 19,056) data whilst the external validation of predictive models was done on the 2010 Medical Expenditure Panel Survey. Using the prediction model, the number of adults who would be detected with prediabetes and/or type 2 diabetes if screened under ADA or USPSTF guidelines were estimated. The ADA screening model was proved more robust.
Meng et al, 2013	Diabetes or prediabetes	12 established risk factors were used.	Logistic regression, ANN, and decision tree	The sample consisted of 1487 (735 volunteered prediabetic/diabetic patients and 752 normal controls) individuals. A standard questionnaire was used to obtain information on risk factors. Three optimal predictive models, namely, a decision tree, a logistic regression model and an ANN having classification accuracies of 77.87%, 76.13% and 73.23% respectively, were constructed.
Yokota et al, 2017	Conversion of prediabetes to diabetes	10 established risk factors and alanine aminotransferase were used	Logistic regression	Aim of the study was to clarify the natural course of prediabetes and develop predictive models for conversion to diabetes. Data of 2105 prediabetic adults from a retrospective longitudinal study were used for building models. The prediction model for incident diabetes had an AUC of 0.80. Moreover, weight reduction was associated with lowered conversion rate.
Morteza et al, 2013	Albuminuria in T2DM	Sex, duration of diabetes, SBP, DBP, GFR, HDL, LDL, TG, HDL/TG ratio, cholesterol, FPG, & Hb1Ac were used as predictors	ANN & conditional logistic regression	Data from a matched case-control study (N = 1104) were used. The two predictive models elucidated markedly different predictors. Authors concluded that the ANN model **complements the current risk factor models** to improve the care of diabetic patients
Kandhasamy et al, 2015	diabetes	8 predictors; number of times pregnant, plasma glucose concentration, DBP, triceps skin fold thickness, 2-hour serum insulin, BMI, diabetes pedigree function, and age, were used.	J48 decision tree, KNN, RF, SVM	Aim of the study was to compare the diabetes prediction performance of four different machine learning algorithms. The algorithms were tested with data samples downloaded from UCI machine learning data repository. Performances of the algorithms were measured in two different contexts, namely, on a dataset with noisy data (before pre-processing) and on a dataset set without noisy data (after pre-processing). The performances were compared in terms of accuracy, sensitivity, and specificity. Before pre-processing, decision tree J48 classifier achieved the highest accuracy of 73.82 %. After pre-processing: both KNN (k=1) and random forest provided 100% accuracy
Jahani & Mahdavi, 2016	Diabetes	Age, fasting blood sugar, diastolic blood pressure, BMI, and change in body weight were used	ANN	The aim of the study was to predict diabetes using clinical and lifestyle characteristics. Memetic algorithms and parameter tuning were used to enhance model accuracy. The sample consisted of 515 individuals; 345 diabetic and 170 non-diabetic. The memetic algorithm improved the accuracy from 88.0% to 93.2%. For the memetic algorithm model, sensitivity, specificity, positive predictive value, negative predictive value, and AUC were 96.2, 95.3, 93.8, 92.4, and 0.958 respectively. Authors argued that the results of this study provide a basis to design a decision support system for risk management and planning of care for individuals at risk of diabetes.

Farran et al, 2013	T2DM, HT, & comorbidities	4 algorithms; KNN, logistic regression, multifactor dimensionality reduction & SVM	Data from a retrospective cohort study were used. Fivefold cross validation was applied to prediction models to obtain generalization accuracy. The sample consisted of 270 172 hospital visitors (of which, 89 858 were diabetic, 58 745 hypertensives and 30 522 comorbid) and included Kuwaiti natives, as well as Asian and Arab expatriates. Optimal classification accuracies of >85% (for diabetes) and >90% (for hypertension) were achieved using only simple non-laboratory-based parameters. Risk assessment tools based on KNN models assigned 'high' risk to 75% of diabetic patients and to 94% of hypertensive patients. Only 5% of diabetic patients were assigned 'low' risk. Two-stage aggregate classification models and risk assessment tools built combining both the component models on diabetes (or on hypertension), performed better than individual models. Authors concluded that these tools aid in the preliminary non-intrusive assessment of the population. Ethnicity was a significant predictor. The need for building risk assessment tools using regional data was highlighted by demonstrating the applicability of the American Diabetes Association online calculator on data from Kuwait.

Abbreviations: CART-classification and regression tree, NHANES-National Health and Nutrition Examination Survey, AUC-area under the receiver operating characteristic curve, US-United States, ANN-artificial neural network, SVM-support vector machine, ICD-international classification of diseases, ADA-American Diabetes Association, USPSTF-United States Preventive Services Task Force, T2DM-type 2 diabetes mellitus, SBP-systolic blood pressure, DBP-diastolic blood pressure, GFR- glomerular filtration rate, HDL-high density lipoproteins, LDL-low density lipoproteins, TG-triglycerides, FPG-fasting plasma glucose, Hb1Ac-glycated hemoglobin, BMI-body mass index, KNN-k nearest neighbors, RF-random forest, UCI-University of California, Irvine, HT-hypertension

References

Anderson, J.P., Parikh, J.R., Shenfeld, D.K., Ivanov, V., Marks, C., Church, B.W., Laramie, J.M., Mardekian, J., Piper, B.A., Willke, R.J. and Rublee, D.A., 2016. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*, *10*(1), pp.6-18.

Baan, C.A., Ruige, J.B., Stolk, R.P., Witteman, J.C., Dekker, J.M., Heine, R.J. and Feskens, E.J., 1999. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. *Diabetes care*, *22*(2), pp.213-219.

Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.H., Kang, E.S. and Kim, D.W., 2014. Screening for prediabetes using machine learning models. *Computational and mathematical methods in medicine*, *2014*.

Dall, T.M., Narayan, K.V., Gillespie, K.B., Gallo, P.D., Blanchard, T.D., Solcan, M., O'Grady, M. and Quick, W.W., 2014. Detecting type 2 diabetes and prediabetes among asymptomatic adults in the United States: modeling American Diabetes Association versus US Preventive Services Task Force diabetes screening guidelines. *Population health metrics*, *12*(1), p.12.

Farran, B., Channanath, A.M., Behbehani, K. and Thanaraj, T.A., 2013. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ open*, *3*(5), p.e002457.

Heikes, K.E., Eddy, D.M., Arondekar, B. and Schlessinger, L., 2008. Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. *Diabetes care*, *31*(5), pp.1040-1045.

Jahani, M. and Mahdavi, M., 2016. Comparison of predictive models for the early diagnosis of diabetes. *Healthcare informatics research*, *22*(2), pp.95-100.

Kandhasamy, J.P. and Balamurali, S., 2015. Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, pp.45-51.

Lee, Y.H., Bang, H., Kim, H.C., Kim, H.M., Park, S.W. and Kim, D.J., 2012. A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores. *Diabetes care*, *35*(8), pp.1723-1730.

Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q. and Liu, Q., 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), pp.93-99.

Morteza, A., Nakhjavani, M., Asgarani, F., Carvalho, F.L., Karimi, R. and Esteghamati, A., 2013. Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression. *Translational research*, *161*(5), pp.397-405.

Ogunyemi, O. and Kermah, D., 2015. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 983). American Medical Informatics Association.

Yokota, N., Miyakoshi, T., Sato, Y., Nakasone, Y., Yamashita, K., Imai, T., Hirabayashi, K., Koike, H., Yamauchi, K. and Aizawa, T., 2017. Predictive models for conversion of prediabetes to diabetes. *Journal of diabetes and its complications*, *31*(8), pp.1266-1271.

Table 2: Key R statistical packages and their functions used for variable importance estimation, resampling methods and other analytical steps employed in the study

Algorithm	R statistical packages and the functions used	Variable importance estimation method
Logistic regression	The "glm" function in the stats package was used for logistic regression modelling (R Core Team, 2018). Ten-fold cross-validation was applied using the "trainControl" function of the "caret" package (Kuhn et al, 2017). The "ovun.sample" function in the ROSE (Random Oversampling Examples) package was used to create under- and over-sampled datasets while the "ROSE" function of the same package was used to create a re-sampled dataset as per ROSE algorithm (Lunardon et al, 2014). The "SMOTE" function in the "DMwR" (Data Mining with R) package was used to create a re-sampled dataset as per the SMOTE (Synthetic Minority Over-Sampling Technique) algorithm (Torgo, 2016). The AUC of prediction models were calculated using "ROCR" package (Sing et al, 2005). The "OptimalCutpoints" package was used to determine cut-off values that maximize the Youden index for performance evaluation metrics of predictive models (López-Ratón et al, 2014).	The OR, CI and p-values were calculated using default functions available in the stats package for generalized linear models.
Artificial neural network	"Model Averaged Neural Network" ("'avNNet') modelling method of the "nnet" package (Venables & Ripley, 2002) was applied through the "train" wrapper function of the "caret" package (Kuhn et al, 2017). The "findCorrelation" function of the "caret" (Classification and Regression Training") package (Kuhn et al, 2017) was used to take a correlation matrix and determine the column numbers that should be removed to keep all pair-wise correlations below a threshold level of 0.75. A specific candidate set of models were created using the "expand.grid" function of the "base" package (R Core Team, 2018) specifying the ranges for specific parameters, namely, size (number of units in hidden layer), decay (regularization parameter to avoid over-fitting) and bag (bagging). Five-fold cross-validation was applied using the "trainControl" function of the "caret" package (Kuhn et al, 2017). The "train" function of the "caret" package (Kuhn et al, 2017) was used to specify model parameters which included	The default function in the "caret" package for estimating the variable importance values of neural network models was used. The default method is based on Gevrey et al (2003), which uses combinations of the absolute values of the weights. For classification models, the class-specific importance values will be the same. All measures of importance are scaled to have a maximum value of 100 by default, but since the scale argument was set to FALSE, unscaled figures are given.

	commands to aggregate neural networks using model averaging, pre-specified weight, and decay tuning parameters, five-fold cross-validation, automatic standardization of data prior to modeling and prediction, the maximum allowable number of weights, and the maximum number of iterations. The"pROC" package was used to calculate the AUC of prediction models (Robin et al, 2011). The "OptimalCutpoints" package was used to determine cut-off values that maximize the Youden index for performance evaluation metrics of predictive models (López-Ratón et al, 2014). The "trainControl" function of the "caret" package was used for resampling; its "sampling" argument was specified as "down", "up". "smote" and "rose" to create under-sampled, over-sampled, "SMOTE" and "ROSE" resampled datasets respectively (Kuhn et al, 2017). The "dummyVars" function of the "caret" package was used to discretize categorical variables (Kuhn et al, 2017).	
Random forest	Default functions for random forests modelling in "randomForest" package (Liaw & Wiener, 2002) along with the "train" wrapper function of the "caret" package (Kuhn et al, 2017) was used. Default values used: "mtry" (number of variables randomly sampled as candidates at each split) parameter's default value for classification is (sqrt(p) where p is the number of variables in the dataset, "ntree" (number of trees to grow) which by default is set to 500, and "nodesize" (minimum size of terminal nodes), the default value for which in a classification model is 1. Ten-fold cross-validation was applied using the "trainControl" function of the "caret" package (Kuhn et al, 2017). The "train" function of the "caret" package (Kuhn et al, 2017) was used to specify model parameters which included commands for ten-fold cross-validation, automatic standardization of data prior to modeling and prediction, and default values for other arguments within the random forests algorithm The"pROC" package was used to calculate the AUC of prediction models (Robin et al, 2011). The "OptimalCutpoints" package was used to determine cut-off values that maximize the Youden index for performance evaluation metrics of predictive models (López-Ratón et al,	By default, mean decrease in prediction accuracy after a variable is permuted is presented as variable importance estimates. It is computed from permuting "out-of-bag" (OOB) data: For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences.

	2014). The “trainControl” function of the “caret” package was used for resampling; its “sampling” argument was specified as “down”, “up”. “smote” and “rose” to create under-sampled, over-sampled, “SMOTE” and “ROSE” resampled datasets respectively (Kuhn et al, 2017).	
Gradient boosting	Default functions for gradient boosting classification tree model in “xgboost” (Extreme Gradient Boosting) package (Chen et al, 2017) along with the “train” wrapper function of the “caret” package (Kuhn et al, 2017) were used; “nrounds” (number of boosting iterations) default = 100, “eta” (step size shrinkage used in update to prevents overfitting a.k.a. learning rate) default = 0.3, “gamma” (minimum loss reduction required to make a further partition on a leaf node of the tree a.k.a. minimum split loss) default = 0, “max_depth” (maximum depth of a tree) default = 6, “min_child_weight” (minimum sum of instance weight (hessian) needed in a child) default=1, “max_delta_step” (Maximum delta step we allow each tree’s weight estimation to be) default=0, “subsample” (subsample ratio of the training instance) default=1, “colsample_bytree” (subsample ratio of columns when constructing each tree) default=1, “colsample_bylevel” (subsample ratio of columns for each split, in each level) default=1, and “lambda” (L2 regularization term on weights) default=1, and “alpha” (L1 regularization term on weights) default = 0. Ten-fold cross-validation was applied using the “trainControl” function of the “caret” package (Kuhn et al, 2017). The “train” function of the “caret” package (Kuhn et al, 2017) was used to specify model parameters which included commands for ten-fold cross-validation, automatic standardization of data prior to modeling and prediction, and default values for other arguments within the gradient boosting classification "xgbTree" algorithm described above. The”pROC” package was used to calculate the AUC of prediction models (Robin et al, 2011). The “OptimalCutpoints” package was used to determine cut-off values that maximize the Youden index for performance evaluation metrics of predictive models (López-Ratón et al, 2014). The “trainControl” function of the “caret” package was used for resampling; its “sampling” argument was specified as “down”, “up”. “smote” and “rose” to create under-sampled, over-sampled, “SMOTE” and “ROSE” resampled datasets	The default function in the “caret” package for estimating the variable importance values of gradient boosting models was used. This method uses the same approach as a single tree (i.e. reduction in the loss function attributed to each variable at each split is tabulated, summed over each node, and totaled.) but sums the importance estimates over each boosting iteration. Estimates are scaled relative to the best performing variable. The variable with the highest sum of improvements is scored 100, and all other variables will have lower scores ranging downwards toward zero.

	respectively (Kuhn et al, 2017). The "dummyVars" function of the "caret" package was used to discretize categorical variables (Kuhn et al, 2017).	
Bagged CART	The bagged CART modelling facility included in the "caret" package (Kuhn et al, 2017) via a wrapper for "train" function was used which is implemented via packages "ipred" (Peters & Hothorn, 2017), "plyr" (Wickham, 2011) and "e1071" (Meyer et al, 2017). The "caret" package provides no tuning parameters for this model. The default parameters were therefore used including those as specified in "bagging" and "rpart.control" functions (Kuhn et al, 2017). Ten-fold cross-validation was applied using the "trainControl" function of the "caret" package (Kuhn et al, 2017). The "train" function of the "caret" package (Kuhn et al, 2017) was used to specify the model which included commands for ten-fold cross-validation, automatic standardization of data prior to modeling and prediction, and default values for other arguments within the bagged CART classification (method = "treebag") The"pROC" package was used to calculate the AUC of prediction models (Robin et al, 2011). The "OptimalCutpoints" package was used to determine cut-off values that maximize the Youden index for performance evaluation metrics of predictive models (López-Ratón et al, 2014). The "trainControl" function of the "caret" package was used for resampling; its "sampling" argument was specified as "down", "up". "smote" and "rose" to create under-sampled, over-sampled, "SMOTE" and "ROSE" resampled datasets respectively (Kuhn et al, 2017).	The default function in the "caret" package for estimating the variable importance values of bagged CART models was used. The same methodology as a single tree (i.e. The reduction in the loss function attributed to each variable at each split is tabulated and the sum is returned.) is applied to all bootstrapped trees and the total importance is returned. Estimates are scaled relative to the best performing variable. The variable with the highest sum of improvements is scored 100, and all other variables will have lower scores ranging downwards toward zero.

References

Chen, T., He, T., Benesty, M., Khotilovich, V. and Tang, Y., 2017. Xgboost: extreme gradient boosting. *R package version 0.6.4.6.* https://github.com/dmlc/xgboost

Gevrey, M., Dimopoulos, I. and Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), pp.249-264.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. and Engelhardt, A., 2017. Caret: classification and regression training. 2017. *R package version*, *4*. https://CRAN.Rproject.org/package=caret

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.

López-Ratón, M., Rodríguez-Álvarez, M.X., Cadarso-Suárez, C. and Gude-Sampedro, F., 2014. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, *61*(8), pp.1-36. http://www.jstatsoft.org/v61/i08/

Lunardon, N., Menardi, G. and Torelli, N., 2014. ROSE: A Package for Binary Imbalanced Learning. *R Journal*, *6*(1).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F., 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2017. *R package version 1.6-8*. https://CRAN.R-project.org/package=e1071

Peters, A. and Hothorn, T., 2017. ipred: Improved predictors. R package version 0.9-6. https://CRAN.R-project.org/package=ipred

R Core Team., 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C. and Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), p.77. http://www.biomedcentral.com/1471-2105/12/77

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T., 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), pp.3940-3941. http://rocr.bioinf.mpi-sb.mpg.de

Torgo, L., 2016. *Data mining with R: learning with case studies*. CRC press. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

Venables, W.N., Ripley. BD, 2002. Modern applied statistics with S. *New York: Springer Science & Business Media*, *200*, pp.183-206.

Wickham, H., 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), pp.1-29. http://www.jstatsoft.org/v40/i01/

Objective 1

Table 3: Variables pre-selected and included in feature selection process for the objective 1 and the rationale/evidence for their inclusion

	Variable	Explanation and/or its construction and coding	Association with prediabetes and/or diabetes/rationale
Questionnaire-based/self-reported			
1	Diagnosed DM risk	Question from the NHANES questionnaire: "Have you ever been told by a doctor or other health professional that you have health conditions or a medical or family history that increases your risk for diabetes?" Responses were coded as: 1 = no, 2 = yes	A prospective cohort study (Lyssenko et al, 2005) found that a positive family history of diabetes was a risk factor for incident diabetes. Perreault et al (2012) found that regression from pre-diabetes to normal glucose regulation is associated with long-term reduction in diabetes risk. In the present study, it was assumed that being diagnosed as having a high risk for developing diabetes is a potential predictor of prediabetes.
2	Self-perceived DM risk	Question from the NHANES questionnaire: "Do you feel you could be at risk for diabetes or prediabetes?" Responses were coded as: 1 = no, 2 = yes	A study by Portnoy et al (2014) revealed that worry; the most common measure of affective perceptions of vulnerability, was a significant predictor of protective behavioral intentions among T2DM patients and that cognitive and affective perceptions interact. It was assumed that self-perceived DM risk may be associated with prediabetes.
3	Blood test for DM	Question from the NHANES questionnaire: "Have you had a blood test for high blood sugar or diabetes within the past three years?" Responses were coded as: 1 = no, 2 = yes	Portnoy et al (2014) found the impact of behavioral factors on diabetes. It was assumed that conducting blood tests for diabetes indicates higher self-perceived risk of the disease and maybe associated with prediabetes.
4	Gender	Gender. Responses were coded as: 1 = female, 2 = male	A cross-sectional study on the determinants of HRQOL among T2DM patients reported that compared with men, scores of all QOL domains were lower in women (Shamshirgaran et al, 2014). Gender differences in risk factor control and treatment profile in diabetes (Nilsson et al, 2004; Perreault et al, 2008), gender differences in prediabetes and insulin resistance among obese children (Tester et al, 2013), as well as sex differences in endothelial function markers before conversion to prediabetes (Donahue et al, 2007) have been reported.
5	**Age**	**Question from the NHANES questionnaire: "Age in years of the participant at the time of screening." Individuals 80 and over are top coded at 80 years of age.**	**Both CDC prediabetes screening test (CDC prediabetes screening test, 2018) and the ADA diabetes/prediabetes screening test (ADA diabetes screening test, 2018) allocate the highest risk scores to advanced age groups. Age is a risk factor of prevalent prediabetes as per Hilawe et a (2016). Co-morbid conditions most prevalent among older people such as frailty and sarcopenia are predictors of adverse health outcomes in diabetic patients (Liccini et al, 2016).**
6	Race	The variable was	A population-based cross-sectional study by

		dichotomized as 1 = (non-Hispanic) White, 2 = other	Mainous et al (2014) on the prevalence and predictors of prediabetes in England from 2003 to 2011 found marked socio-economic disparities in the disease prevalence. It was assumed in the present study that race could be a potential predictor of prediabetes among the US population.
7	Served in armed forces	Question from the NHANES questionnaire: "Have you ever served on active duty in the U.S. Armed Forces, military Reserves, or National Guard?" Data were collected from both males and females 17 - 80 years of age. Therefore, missing data among those >= 17 years were multiply-imputed while those among < 17 years were coded as "no". Assigned codes were: 1 = no, 2 = yes.	Work characteristics have been reported to be predictors of diabetes incidence among apparently healthy employees (Toker et al, 2012). It was assumed that occupation in the armed forces may be a predictor of prediabetes in the present study. Also see above about the study by Mainous et al (2014).
8	Country of birth	Question from the NHANES questionnaire: "In what country were you born?" Responses were coded as: 1 = USA, 2 = other	Assumed as a potential socio-economic determinant of prediabetes.
9	**Citizenship**	**Question from the NHANES questionnaire: "Are you a citizen of the United States?" Responses were coded as 1 = yes, 2 = no.**	**Socio-economic deprivation was a predictor of prediabetes as per Mainous et al (2014). Significant differences in prevalence of prediabetes by demographic factors such as race and gender were observed by Zhang et al (2014). Hence a possible impact of citizenship on prediabetes was speculated in this study.**
10	**Marital status**	**The variable was dichotomized as 1 = unmarried/other, 2 = married/ living with partner**	**Shamshirgaran et al (2014) revealed that the quality of life among diabetic individuals was influenced by marital status. Hence a possible influence on prediabetes was speculated in this study.**
11	Household size	Total number of people in the household	Assumed as a potential socio-economic predictor of the disease.
12	Family size	Total number of people in the family	Family context was found to predict T2DM management (Chesla et al, 2003)
13	Household income	Total annual household income (reported as a range value in dollars)	Kohler & Soldo (2005) revealed a potentially life-long impact of socio-economic factors on diabetes.
14	Family income	Total annual family income (reported as a range value in dollars)	Kohler & Soldo (2005); see no:13 above
15	**Income-poverty ratio**	**Ratio of family income to poverty. (range of values 0 to 5)**	**Kohler & Soldo (2005) revealed that socio-economic factors were associated with diabetes. Tsenkova et al (2014) found a lifelong impact of socioeconomic disadvantage on prediabetes. Hence, this variable was included as a proxy indicator of socio-economic status in the present study.**
16	Alcohol use	Question from the NHANES questionnaire: "In any one	An association between alcohol consumption and risk of pre-diabetes and type 2 diabetes

		year, have you had at least 12 drinks of any type of alcoholic beverage?" Responses were coded as 1 = no, 2 = yes.	development was reported in a Swedish population (Cullmann et al, 2012). Alcohol consumption, heavy consumption in particular, was an independent risk factor for the development of prediabetes, but not for diabetes in middle-aged and elderly Chinese men (Zhang, S. et al, 2016).
17	**Diagnosed hypertension**	**Question from the NHANES questionnaire: "Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?" Responses were coded as 1 = no, 2 = yes.**	**Studies by Okwechime et al (2015) and Soewondo & Pramono (2011) revealed a positive association between prediabetes and hypertension. Prediabetes is considered an underlying etiology of metabolic syndrome and hypertension in turn is an integral feature of the metabolic syndrome (Mayans, 2015).**
18	Self-measured BP	Question from the NHANES questionnaire: "Did you take your blood pressure at home during the last 12 months?" Responses were coded as 1 = no, 2 = yes.	See no:17 above.
19	Advised, self-measured BP	Question from the NHANES questionnaire: "Did a doctor or other health professional tell you to take your blood pressure at home?" Responses were coded as 1 = no, 2 = yes	See no:17 above.
20	Diagnosed hypercholesterolemia	Question from the NHANES questionnaire: "Have you ever been told by a doctor or other health professional that your blood cholesterol level was high?" Responses were coded as 1 = no, 2 = yes.	Okwechime et al (2015) revealed a positive association between hypercholesterolemia and prediabetes. Mamtani et al (2016) argued that disturbed lipid metabolism is a hallmark of diabetes and developed and validated a cost-effective plasma lipidomic risk score (LRS) as a biomarker of future T2DM
21	Gastrointestinal disease	Question from the NHANES questionnaire: "Did you have a stomach or intestinal illness with vomiting or diarrhea that started during those 30 days?" Responses were coded as 1 = no, 2 = yes.	A study by Gagnon et al (2017) was premised on the rationale that diabetic individuals are at heightened risk for developing comorbid eating disorders than non-diabetic counterparts and identified predictors of comorbid eating disorders in patients with T1DM or T2DM, which included the type of diabetes, body mass indexes, coping styles and depressive symptoms. An increased frequency of prediabetes in patients with irritable bowel syndrome has also been reported (Gulcan et al, 2009).
22	Self-reported dietary health	Question from the NHANES questionnaire: "How healthy is the diet?" Responses were coded as 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.	Dietary habits and leisure-time physical activity were related to adiposity, dyslipidemia, and incident dysglycemia in the pathobiology of prediabetes in a biracial cohort study (Boucher et al, 2015). Nouwen et al (2011) revealed longitudinal motivational predictors of dietary self-care and diabetes control in adults with newly-diagnosed T2DM. An interventional study found that intensive lifestyle intervention can result in long-term reductions in total energy intake. Initial success in achieving reductions in fat and energy intake and success in attaining activity goals appear to predict long-term

			success at maintaining changes (Davis et al, 2013).
23	Milk consumption	Question from the NHANES questionnaire: "In the past 30 days, how often did you have milk to drink or on your cereal?" Responses were dichotomized as 1 = never or rarely, 2 = more frequently	Early cow's milk exposure was found a potential determinant of subsequent T1DM that might increase the risk nearly 1.5 times (Gerstein, 1994). Among adults, the consumption of milk and dairy products was associated with a markedly reduced prevalence of the metabolic syndrome which is characterized by incident diabetes among other indicators (Elwood et al, 2007).
24	**Food security**	**Household food security category for last 12 months. Responses were recoded as: 1 = very low food security, 2 = low food security, 3 = marginal food security 4 = full food security and included in the models as a numeric variable.**	**A study done on the NHANES 1999-2002 data found that food insecurity was a risk factor for diabetes. Authors suggested that increased consumption of inexpensive, high-caloric food alternatives, among adults with food insecurity could play a part in the observed relationship (Seligman et al, 2007). Mainous et al (2014) revealed the impact of socio-economic factors on prediabetes. Food security was added as a proxy indicator of socio-economic status and access to nutritious food.**
25	Health insurance	Question from the NHANES questionnaire: "Are you covered by health insurance or some other kind of health care plan?" Responses were coded as: 1 = no, 2 = yes.	The association between health insurance coverage and diabetes care was revealed by a study that used data from the 2000 Behavioral Risk Factor Surveillance System (Nelson et al, 2005). Mainous et al (2014) revealed the impact of socio-economic factors on prediabetes and the access to health insurance is presumably linked to socio-economic status and thus was included in the present study.
26	**Hepatitis B**	**Question from the NHANES questionnaire: "Has a doctor or other health professional ever told you that you have hepatitis B?" Responses were coded as: 1 = no, 2 = yes.**	**Gisi et al (2017) found a positive association between diabetes and hepatitis B. Khalili et al (2015) also found a similar positive association and suggested that the resulting diabetes could be attributable to known metabolic risk factors and liver damage, as determined by elevated ALT levels.**
27	**Hepatitis C**	**Question from the NHANES questionnaire: "Has a doctor or other health professional ever told you that you have hepatitis C?" Responses were coded as: 1 = no, 2 = yes.**	**Burman et al (2015), Ali et al (2014) and Mukhtar et al (2012) found hepatitis C was a strong predictor of prediabetes. Howard et al (2003) and Wang et al (2003) revealed a similar positive association with diabetes.**
28	Self-rated general health	Question from the NHANES questionnaire: "Would you say your health in general is…?" Assigned response codes were: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.	An association between self-rated health and diabetes of long duration was reported by Klein et al (1998). Another study found that self-rated health is a significant predictor of mortality in people with late-onset diabetes but not in those with early onset diabetes when physical health status is controlled (Dasbach et al, 1994).
29	Self-rated health trend	Question from the NHANES questionnaire: "Compared with 12 months ago, would you say your health is now…?" Responses were	See no:28 above

		recoded as 1 = better, 2 = about the same, 3 = worse	
30	Healthcare availability	Question from the NHANES questionnaire: "Is there a place that you usually go when you are sick, or you need advice about your health?" Responses were dichotomized as 1 = no, 2 = yes.	A study found that rates of health care access and utilization, screening for diabetes complications, and treatment of hyperglycemia, hypertension, and dyslipidemia in type 2 diabetes are high but health status and outcomes are unsatisfactory. Suggested reasons for this discrepancy were the intractability of diabetes to current therapies, patient self-care practices, physician medical care practices, and characteristics of U.S. health care systems (Harris, 2000). A study on healthcare seeking among African Americans recently identified with prediabetes or early diabetes reinforced the importance of preventive healthcare and sustainable changes in diet and physical activity and concluded that physicians need better tools for motivating and supporting their patients to adopt behaviors that can reduce diabetes risk (Harris & Chew, 2014)
31	Healthcare use frequency	Question from the NHANES questionnaire: "During the past 12 months, how many times have you seen a doctor or other health care professional about your health at a doctor's office, a clinic, or some other place?" (Range from 0 to 16/more)	See no:30 above
32	Mental healthcare use	Question from the NHANES questionnaire: "During the past 12 months, have you seen or talked to a mental health professional such as a psychologist, psychiatrist, psychiatric nurse, or clinical social worker about your health?" Responses were coded 1 = no, 2 = yes	A meta-analysis found that the presence of diabetes doubled the odds of comorbid depression (Anderson et al, 2001). A solid body of evidence exists for the relationship between mood disorders and diabetes mellitus (Musselman et al, 2003). Mental healthcare use was considered a proxy indicator of a mental health disorder and was included in the study.
33	Number of rooms in household	Question from the NHANES questionnaire: "How many rooms are in this home?"	Mainous et al (2014) revealed the impact of socio-economic factors on prediabetes. Assumed as a potential socio-economic determinant of prediabetes.
34	Type of house	Question from the NHANES questionnaire: "Is this mobile home/house/apartment owned, being bought, rented, or occupied by some other arrangement by you/you or someone else in your family?" Responses were dichotomized as 1 = owned, 2 = rented or other	Poor housing conditions was an independent risk factor of incident diabetes in urban, middle-aged African Americans (Schootman et al, 2007). Ownership of a house may also be a potential socio-economic variable.
35	Hepatitis A vaccination	Question from the NHANES questionnaire: "Have you ever received hepatitis A vaccine?"	An association of hepatitis A virus infection with T1DM has been repoted (Hasosah et al, 2016). Another case report on hepatitis A-induced diabetes

		Responses were recoded as 1 = no, 2 = yes; fully or partially	was reported by Vesely et al (1999). Hepatitis A vaccination exerts a protective effect against the disease and hence was assumed to be associated with diabetes.
36	Hepatitis B vaccination	Question from the NHANES questionnaire: "Have you ever received the 3-dose series of the hepatitis B vaccine?" Responses were recoded as 1 = no, 2 = yes; fully or partially	See no:26 above. Since hepatitis B was found to be associated with diabetes, it was assumed that hepatitis B vaccination may be related to diabetes.
37	Monthly family income	Monthly family income (reported as a range of values $0-$8400/above; coded 1-12)	Assumed as a potential socio-economic determinant of prediabetes.
38	Poverty index	Family monthly poverty level index; a ratio of monthly family income to the HHS poverty guidelines specific to family size (range of values 0 – 5)	Assumed as a potential socio-economic determinant of prediabetes.
39	Poverty category	Family monthly poverty level category (1 = Monthly poverty level index <= 1.30, 2 = 1.30 < Monthly poverty level index <= 1.85, 3 = Monthly poverty level index > 1.85)	Assumed as a potential socio-economic determinant of prediabetes.
40	Diagnosed kidney disease	Question from the NHANES questionnaire: "Have you ever been told by a doctor or other health professional that you had weak or failing kidneys?" Responses were coded as 1 = no, 2 = yes.	Tabá et al (2012) reported that nephropathy and CKD are associated with prediabetes. A study that used NHANES 1999-2006 data found that CKD prevalence was high among people with undiagnosed diabetes and prediabetes (Plantinga et al, 2010). A prospective cohort study showed an independent role of prediabetes in development of hyperfiltration and albuminuria indicating prediabetes might be a target for early treatment to prevent chronic kidney disease in chronic hyperglycemia (Melsom et al, 2016).
41	Kidney stones	Question from the NHANES questionnaire: "Have you ever had kidney stones?" Responses were coded as 1 = no, 2 = yes.	A study done using the NHANES III 1988-1994 data reported that metabolic syndrome (traits of which include diabetes) were associated with a self-reported history of kidney stones (West et al, 2008). Chu et al (2017) reported an association of uric acid calculi with prediabetic and diabetic states.
42	Urinary leakage frequency	Question from the NHANES questionnaire: "How often do you have urinary leakage? Would you say…?" Responses were coded as 1 = never, 2 = less than once a month, 3 = a few times a month, 4 = a few times a week, 5 = every day and/or night	A prospective cohort study revealed an association between T2DM and urinary incontinence in women (Lifford et al,2005). Both polyuria and polydipsia were found to be independent predictors of prediabetes in a study done by Charfen et al (2009).
43	Nightly urinating frequency	Question from the NHANES questionnaire: "During the past 30 days, how many times per night did you most	Nocturia was found to be an independent predictor of diabetes (FitzGerald et al, 2007; Yoshimura et al, 2004)

		typically get up to urinate, from the time you went to bed at night until the time you got up in the morning?" (0 to 5/more; coded accordingly 0 - 5)	
44	Diagnosed asthma	Question from the NHANES questionnaire: "Has a doctor or other health professional ever told you that you have asthma?" Responses were coded as 1 = no, 2 = yes	A strong positive association between the occurrence of T1DM and symptoms of asthma at the population level in Europe and elsewhere was reported (Stene & Nafstad, 2001).
45	Diagnosed anemia	Question from the NHANES questionnaire: "Were you taking treatment for anemia during past 3 months. Responses were coded as 1 = no, 2 = yes.	Anemia was associated with diabetes particularly in diabetic individuals with albuminuria or reduced renal function (Thomas et al, 2003). Davis et al (2017) showed that anemia was positively associated with diabetic retinopathy
46	Diagnosed psoriasis	Question from the NHANES questionnaire: "Have you ever been told you have psoriasis?" Responses were coded as 1 = no, 2 = yes.	Several studies revealed that psoriasis was associated with diabetes (Shapiro et al, 2007; Qureshi et al, 2009; Armstrong et al, 2013). A similar association between psoriasis and insulin resistance was reported by Pereira et al (2011).
47	Diagnosed celiac disease	Question from the NHANES questionnaire: "Have you ever been told you have celiac disease?" Responses were coded as 1 = no, 2 = yes.	An association between celiac disease and T1DM has been reported (Ludvigsson et al, 2006). Also, an increased prevalence of diabetes-related autoantibodies was reported in celiac disease (Rapoport et al, 1996).
48	Gluten free diet	Question from the NHANES questionnaire: "Are you on a gluten-free diet?" Responses were coded as 1 = no, 2 = yes.	An experimental study found that gluten-free diet both delayed and to a large extent prevented diabetes in NOD mice that have never been exposed to gluten (Funda et al, 1999). Gluten-free diet is indicated for celiac disease patients and see no:47 above regarding its relationship with T1DM.
49	Diagnosed arthritis	Question from the NHANES questionnaire: "Did the doctor ever say you have arthritis?" Responses were coded as 1 = no, 2 = yes.	A positive association between arthritis and prediabetes was reported (Okwechime et al, 2015). A retrospective study at a rheumatology clinic found a higher incidence of prediabetes in patients with rheumatic diseases than in patients with other diseases and recommended healthcare guidance for these high-risk patients to prevent metabolic syndrome (Origuchi et al, 2011).
50	Diagnosed gout	Question from the NHANES questionnaire: "Did doctor ever tell you that you had gout?" Responses were coded as 1 = no, 2 = yes.	a high co-prevalence of gout, diabetes and CVD in the adult population was reported in a study done by Winnard et al (2013). Another study revealed that men with gout are at a higher future risk of T2DM independent of other known risk factors (Choi et al, 2008).
51	Diagnosed congestive heart failure	Question from the NHANES questionnaire: "Were you ever told you had congestive heart failure?" Responses were coded as 1 = no, 2 = yes.	Parildar et al (2013) revealed that prediabetes was positively associated with hypertension and hyperlipidemia. A study of cardiovascular physiological changes among prediabetic patients revealed that the atrial conduction times, and P wave dispersion on surface electrocardiographic were longer before the development of overt diabetes and the left atrial mechanical functions

			were also impaired secondary to a deterioration in the diastolic functions among them (Gudul et al, 2017) Another study revealed that, in patients with heart failure and reduced ejection fraction, dysglycemia is common and prediabetes is associated with a higher risk of adverse cardiovascular outcomes compared with patients with non-prediabetic individuals (Kristensen et al, 2016).
52	Diagnosed coronary heart disease	Question from the NHANES questionnaire: "Have you ever been told you have coronary heart disease?" Responses were coded as 1 = no, 2 = yes.	Parildar et al (2013) revealed that prediabetes was positively associated with HT and hyperlipidemia. A strong association between cardiovascular risk factors and confirmed prediabetes was reported by Haffner et al (1990).
53	Diagnosed angina	Question from the NHANES questionnaire: "Have you ever been told you have angina/angina pectoris?" Responses were coded as 1 = no, 2 = yes.	An association between diabetes and angina pectoris was reported in a prospective cohort study (Mathenge et al, 2017)
54	Diagnosed heart attack	Question from the NHANES questionnaire: "Have you ever been told you have heart attack?" Responses were coded as 1 = no, 2 = yes.	Prediabetes and newly diagnosed diabetes were found to be highly prevalent in patients with a transient ischemic attack or stroke (Fonville et al, 2013). Parildar et al (2013) showed that prediabetes was positively associated with hypertension and hyperlipidemia. Naas et al (1998) revealed that QT and QTc dispersion are accurate predictors of cardiac death in newly diagnosed non-insulin dependent diabetes
55	Diagnosed stroke	Question from the NHANES questionnaire: "Have you ever been told you have a stroke?" Responses were coded as 1 = no, 2 = yes.	An association between stroke and prediabetes was reported by Lee et al (2012).
56	Diagnosed emphysema	Question from the NHANES questionnaire: "Have you ever been told you have emphysema?" Responses were coded as 1 = no, 2 = yes.	A study using the Framingham Heart Study data revealed a strong adverse effect of glycemic state on the pulmonary function (Walter et al, 2003)
57	Diagnosed thyroid disease	Question from the NHANES questionnaire: "Have you ever been told you have a thyroid problem?" Responses were coded as 1 = no, 2 = yes	Thyroid dysfunction was associated in type 2 diabetic patients in several studies (Radaideh et al, 2004; Perros et al, 1995).
58	Diagnosed chronic bronchitis	Question from the NHANES questionnaire: "Have you ever been told you have chronic bronchitis?" Responses were coded as 1 = no, 2 = yes	See no:56 above.
59	Diagnosed liver disease	Question from the NHANES questionnaire: "Have you ever been told you have any liver condition?" Responses were coded as 1 = no, 2 = yes	A retrospective cohort study conducted among prediabetic patients found that those with non-alcoholic fatty liver disease(NAFLD) had a significantly higher risk of DM than those without NAFLD (Nishi et al, 2015). Liver inflammation was found to be a risk factor for prediabetes in at-risk individuals with and without Hepatitis C infection

			in a study done by Burman et al (2015). The NAFLD independently predicted prediabetes during a 7-year prospective follow-up (Zelber-Sagi et al, 2013)
60	Diagnosed COPD	Question from the NHANES questionnaire: "Have you ever been told you have COPD?" Responses were coded as 1 = no, 2 = yes	An association between COPD and diabetes has been reported (Yamane et al, 2013; Stojkovikj et al, 2016)
61	Diagnosed jaundice	Question from the NHANES questionnaire: "Have you ever been told you have jaundice?" Responses were coded as 1 = no, 2 = yes	See no:59 above. A study by Wang, J. et al (2017) revealed that serum bilirubin concentrations were positively associated with the risk of incident T2DM in middle-aged and elderly adults
62	Diagnosed cancer	Question from the NHANES questionnaire: "Have you ever been told you have cancer or malignancy?" Responses were coded as 1 = no, 2 = yes	A study by Huang et al (2014) revealed that prediabetes was positively associated with different types of cancer as well as overall cancer. Also, an association between prediabetes and pancreatic cancer was revealed by a meta-analysis (Fu et al., 2016)
63	Familial heart attack	Question from the NHANES questionnaire: "Have any of your close relatives ever had heart attack?" Responses were coded as 1 = no, 2 = yes	Since cardiovascular disease and complications often co-exist with both prediabetes and diabetes, a shared risk factor profile at the genetic level for both conditions was assumed. A predictive model based on genomic data including several novel loci and single nucleotide polymorphisms (SNPs) was associated with the risk of developing T2DM (Lee et al, 2011). Another genetic association study suggested that the CDKN2A-rs10811661 polymorphism, waist-hip ratio, systolic blood pressure, and dyslipidemia were significantly associated with the increased risk of prediabetes in a Vietnamese population (Binh et al, 2015). A cross-sectional analysis of genetic data demonstrated that at least 6 out of 41 genetic variants characteristic of individuals with T2DM may also be associated with prediabetes. Accumulation of these risk alleles may markedly increase the risk for prediabetes (Zyriax et al, 2013).
64	Familial asthma	Question from the NHANES questionnaire: "Have any of your close relatives ever had asthma?" Responses were coded as 1 = no, 2 = yes	See no:44 and no:63 above. A familial aggregation of the two conditions was speculated.
65	Familial diabetes	Question from the NHANES questionnaire: "Have any of your close relatives had diabetes?" Responses were coded as 1 = no, 2 = yes	A community survey in a rural Thai population revealed that having a blood relative with diabetes was a strong predictor of prediabetes (Lorga et al, 2012). See also no:63 above.
66	Advised to lose weight	Question from the NHANES questionnaire: "Has doctor ever told you to lose weight?" Responses were coded as 1 = no, 2 = yes	Prediabetes was positively associated with obesity/overweight (Okwechime et al, 2015)
67	Advised to exercise	Question from the NHANES questionnaire: "Has doctor	Independent and combined effects of exercise training and pharmacotherapy on insulin sensitivity

		told you to exercise?" Responses were coded as 1 = no, 2 = yes	in prediabetic individuals has been demonstrated (Malin et al, 2012). It is reported that pancreatic β-cell function increases in a linear dose-response manner following exercise training in adults with prediabetes (Malin et al, 2013)
68	Advised to reduce salt intake	Question from the NHANES questionnaire: "Has doctor ever told you to reduce salt in diet?" Responses were coded as 1 = no, 2 = yes	Prediabetes was positively associated with HT (Okwechime et al, 2015)
69	Advised to reduce fat intake	Question from the NHANES questionnaire: "Has doctor told you to reduce fat/calories?" Responses were coded as 1 = no, 2 = yes	Gholi et al (2016) revealed that both the risk factors of CVD and body composition parameters were different between the prediabetic and normal groups; total cholesterol and triglyceride were predictors of the risk of prediabetes.
70	Managing weight	Question from the NHANES questionnaire: "Are you now controlling or losing weight?" Responses were coded as 1 = no, 2 = yes	Prediabetes was found to be positively associated with obesity/overweight (Okwechime et al, 2015)
71	Increasing exercise	Question from the NHANES questionnaire: "Are you now increasing exercise?" Responses were coded as 1 = no, 2 = yes	In a biracial cohort comprised of offsprings of parents with T2DM, exercise habits correlated with measures of adiposity and dyslipidemia and the physical activity significantly predicted incident dysglycemia during 5.5 years of follow-up. (Boucher et al, 2015)
72	Reducing salt intake	Question from the NHANES questionnaire: "Are you now reducing salt in diet?" Responses were coded as 1 = no, 2 = yes	A study revealed that high salt intake increased blood pressure and albuminuria in diabetic patients with microalbuminuria and the responses were associated with insulin resistance and increased glomerular pressure leading to the conclusion that insulin resistance contributes to higher salt sensitivity glomerular pressure & albuminuria (Vedovato et al, 2004).
73	Reducing fat intake	Question from the NHANES questionnaire: "Are you now reducing fat in diet?" Responses were coded as 1 = no, 2 = yes	Davis et al (2013) revealed that an intensive lifestyle intervention can result in long-term changes in dietary behavior and the initial success in achieving reductions in fat and caloric intake predicted long-term success at maintaining changes.
74	Duration of sedentary activity	The duration of sedentary activity in minutes	Anjana et al (2015) revealed that physical inactivity was a predictor of dysglycemia. Another study reported that physical inactivity was an independent predictor of incident diabetes (Longo-Mbenza et al, 2010)
75	Duration of watching TV	Duration of watching TV or videos past 30 days in hours. Responses were recoded as: 0 – don't watch TV/less than 1 hour, 1 = 1 hour, 2 = 2 hours, 3 = 3 hours, 4 = 4 hours, 5 = 5 hours or more	See no:74 above
76	Duration of computer use	Duration of using computer past 30 days in hours. Responses were recoded as: 0 – don't use a computer outside	See no:74 above

		of school /less than 1 hour, 1 = 1 hour, 2 = 2 hours, 3 = 3 hours, 4 = 4 hours, 5 = 5 hours or more	
77	Duration of sleep	Question from the NHANES questionnaire: "How much sleep do you get (hours)?" Responses range 2-11 and 12/more hours, coded accordingly	Physical inactivity was a predictor of dysglycemia as per Anjana et al (2015). A cross-sectional study revealed that nearly 25% of patients with T2DM were diagnosed with a sleep disorder and over 75% reported experiencing at least one sleep symptom regularly (Gupta & Wang, 2016). A systematic review and meta-analysis found that sleep disturbances [short (<6 h) and long (>8 h) sleeping time, insomnia, obstructive sleep apnea and abnormal sleep timing] were associated with incident diabetes (Anothaisintawee et al, 2016). A prospective cohort study showed that people with regular sleep disorders, people with short and long sleep duration, but not regular daytime nappers are at increased risk of diabetes. Furthermore, regular sleep disorders were associated with an increased risk of prediabetes (Kowall et al, 2016)
78	Ever smoking	Question from the NHANES questionnaire: "Have you smoked at least 100 cigarettes in life?" Responses were coded as 1 = no, 2 = yes	A birth cohort analysis of the high-risk glycated hemoglobin trajectories established by mid-20s reported that being a smoker at age 26 predicted membership of the least favorable trajectory over the next 12 years (Shearer et al, 2016).
79	Number of smokers in the household	Number of people living in the household that smoke tobacco. Responses ranged from: 0-3/more and were coded accordingly	A population-based cohort study showed that both passive and active smoking are associated with T2DM (Kowall et al, 2010).
80	Past smoking	Question from the NHANES questionnaire: "Have you smoked tobacco last 5 days?" Responses were coded as 1 = no, 2 = yes	A cross-sectional study revealed that smoking is strongly associated with pre-diabetes in young adults with a low burden of smoking exposure and underscored nicotine dependence could be a potential mechanism of this relationship (Aeschbacher et al, 2014).
81	Past smokeless tobacco use	Question from the NHANES questionnaire: "Have you used smokeless tobacco last 5 days?" Responses were coded as 1 = no, 2 = yes	A prospective cohort study showed that high consumption of snus (a form of smokeless tobacco) predicts risk of developing T2D which may be mediated by effects on beta-cell function (Östenson et al, 2012)
82	Past any tobacco use	Question from the NHANES questionnaire: Have you used any tobacco product last 5 days?" Responses were coded as 1 = no, 2 = yes	See no: 78-81 above
Physiological and biochemical variables			
83	Pulse rate	60 sec. pulse (30 sec. pulse * 2)	A study reported that aortic pulse wave velocity is a strong independent predictor of mortality in diabetes and may represent a useful integrated index of vascular status and hence cardiovascular risk (Cruickshank et al, 2002). Another study found that diabetes adversely affects left ventricular structure and function independently of increases in BMI and

			blood pressure (Devereux et al, 2000).
84	Pulse character	Is the pulse regular or irregular? Responses were coded as 1 = no, 2 = yes	Persons with diabetes or borderline glucose intolerance were found to have stiffer arteries than their counterparts with normal glucose tolerance and the decreased elasticity was independent of artery wall thickness by a study done by Salomaa et al (1995).
85	**Body mass index**	**Body Mass Index (kg/m**2)**	**Okwechime et al (2015) found that obesity/overweight was positively associated with prediabetes. Haghighatdoost et al (2017) found a similar positive association with diabetes.**
86	Arm circumference	Arm Circumference (cm)	Prediabetes was positively associated with obesity/overweight in a study done by Okwechime et al (2015). Mid-upper arm circumference is a proxy indicator of overweight/obesity
87	**Waist circumference**	**Waist Circumference (cm)**	**The World Health Organization (2011) recommended the use of waist circumference over arm circumference as a risk indicator of chronic diseases including diabetes. Okwechime et al (2015) found a positive association between obesity/overweight and prediabetes. Haghighatdoost et al (2017) revealed a similar positive association with diabetes. Lorga et al (2012) found a positive association between impaired fasting plasma glucose status, which is a precursor stage of diabetes, and waist circumference. A study by Bassareo et al (2013) revealed that waist circumference was a better risk indicator of hypertension than arm circumference.**
88	Sagittal abdominal diameter	Average Sagittal Abdominal Diameter (cm)	Sagittal abdominal diameter was found to be a strong anthropometric marker of insulin resistance and hyperproinsulinemia in obese men in a study done by Risérus et al (2004)
89	Toxoplasmosis IgG antibodies	Toxoplasmosis IgG antibodies (IU/ml)	A meta-analysis suggested chronic toxoplasmosis as a possible risk factor for T2DM. However, based on random effects model, no statistically significant association was observed between *T. gondii* and T1DM (Majidiani et al, 2016)
90	Toxoplasmosis IgM antibodies	Toxoplasmosis IgM antibodies	See no: 89 above
91	Urinary albumin	Albumin, urine (ug/mL)	A longitudinal study revealed that albuminuria is positively associated with the conversion from prediabetes to diabetes (Wang et al, 2010).
92	Urinary creatinine	Creatinine, urine (mg/dL)	A prospective cohort study revealed that an elevated albumin: creatinine ratio is associated with higher creatinine, insulin resistance, and cytokine levels and lower 25-hydroxy vitamin D levels in prediabetic individuals. Microalbuminuria is associated with decreased reversal to normoglycemia and increased progression to diabetes. Low 25-hydroxy vitamin D may be associated with increased progression to diabetes, perhaps via modulation of the albumin: creatinine ratio (Dutta et al, 2014)
93	Urinary albumin	Albumin creatinine ratio	See 92 above

	creatinine ratio	(mg/g)	
94	HDL	Direct HDL-Cholesterol (mg/dL)	Franks et al (2007) reported that HDL level is inversely associated with diabetes. Low HDL level was a risk factor of diabetes as per Karim et al (2013).
95	Cholesterol	Total Cholesterol (mg/dL)	Prediabetes was found to be positively associated with hypercholesterolemia (Okwechime et al, 2015). Another study reported that prediabetes is positively associated with coronary atherosclerosis (Kurihara et al, 2013)
96	**WBC count**	**White blood cell count (1000 cells/uL)**	**A follow-up study by Twig et al (2013) revealed that high WBC count is an independent risk factor for diabetes at values well within the normal range. A meta-analysis of observational studies by Gkrania-Klotsas et al (2010) revealed a positive association between WBC count and T2DM. Vozarova et al (2002) found in a longitudinal study that high WBC count is associated with worsening insulin sensitivity and predicts the development of T2DM.**
97	Lymphocytes count	Lymphocyte number (1000 cells/uL)	An association between elevated lymphocyte count and T2DM was reported by Gkrania-Klotsas et al (2010) in their meta-analysis
98	**Monocytes count**	**Monocyte number (1000 cells/uL)**	**Bulum et al (2014) reported an inverse association between the monocyte count and the risk of insulin resistance in T1DM. Gkrania-Klotsas et al (2010) revealed a negative association albeit non-significant, between monocytes count and T2DM.**
99	Neutrophils count	Segmented neutrophils num (1000 cell/uL)	An increased neutrophil count was associated with T2DM in a meta-analysis done by Gkrania-Klotsas et al (2010).
100	Eosinophils count	Eosinophils number (1000 cells/uL)	An increased granulocyte count was associated with diabetes (Gkrania-Klotsas et al, 2010)
101	Basophils count	Basophils number (1000 cells/uL)	An increased granulocyte count was associated with diabetes (Gkrania-Klotsas et al, 2010)
102	**RBC count**	**Red blood cell count (million cells/uL)**	**Simmons (2010) found that an increased red cell count is present in diabetes precursor states, namely, pre-diabetes, obesity, and the metabolic syndrome which could be partially explained by an elevated HbA1c level. In contrast, a lowered RBC count was reported in overt or untreated diabetes (Lin et al, 2014; Al Shehri, 2017; Wang et al, 2013) and possible pathophysiological mechanisms were presented.**
103	**Hematocrit**	**Hematocrit (%)**	**Meisinger et al (2014) found prediabetes and other precursor stages of diabetes were positively associated with an elevated hematocrit. Ziaee et al (2017) revealed that hematological indices are markers of vascular complication and glycemic control in T2DM patients. Tulloch-Reid et al (2004) found that an elevated hematocrit is associated with higher risk of incident T2DM, possibly mediated through an association with insulin resistance. Nakanishi et al (2004) and**

			Tamariz et al (2008) reported a linear association of hematocrit level with the risk of incident T2DM.
104	Hemoglobin	Serum hemoglobin (g/dL)	Anemia is positively associated with diabetic retinopathy (Davis et al, 2017). A study on the association between mean arterial blood pressure and blood viscosity in children with T1DM and healthy controls revealed that non-diabetic children compensate for the increase in vascular resistance due to increased blood viscosity (increased emoglobin and hematocrit) while diabetic children do not, probably due to endothelial dysfunction (Salazar Vázquez et al, 2010). Another study reported that elevated blood viscosity and hematocrit were found to be emerging risk factors for insulin resistance and T2DM (Tamariz et al, 2008). A study by Li et al (2017) revealed that high hemoglobin and hematocrit are associated with triglycerides, low-density lipoprotein and high-density lipoprotein and negatively associated with diabetes, in patients at righ Risk of coronary artery disease.
105	Platelet count	Platelet count (1000 cells/uL)	A Japanese study reported that platelet count is independently associated with insulin resistance in non-obese type 2 diabetic patients (Taniguchi et al, 2003)
106	Hepatitis A antibody	Hepatitis A antibody. Responses were dichotomized as 1 = negative, 2 = intermediate/positive	See no: 35 above. Liver inflammation was found to be a risk factor for prediabetes in at-risk Latinos with and without Hepatitis C infection indicating liver inflammation, regardless of the type of hepatitis infection, may trigger prediabetes (Burman et al, 2015)
107	Hepatitis B core antibody	Hepatitis B core antibody. Responses were coded as 1 = negative, 2 = positive	See no:26 and no:36 above.
108	Hepatitis B surface antigen	Hepatitis B surface antigen. Responses were coded as 1 = negative, 2 = positive	See no: 26 and no: 36 above
109	Hepatitis D	Hepatitis D (anti-HDV). Responses were coded as 1 = negative, 2 = positive	See no: 106 above
110	Hepatitis B surface antibody	Hepatitis B Surface Antibody. Responses were coded as 1 = negative, 2 = positive	See no: 26 and no:36 above
111	Hepatitis E IgG	Hepatitis E IgG (anti-HEV). Responses were coded as 1 = negative, 2 = positive	See no: 106 above. A study reported that diabetic patients are prone to develop severe hepatitis and liver failure due to hepatitis virus infection (Singh et al, 2013).
112	Hepatitis E IgM	Hepatitis E IgM (anti-HEV). Responses were coded as 1 = negative, 2 = positive	See no: 111 above
113	Alkaline phosphatase	Serum alkaline phosphatase (IU/L)	Elevation of serum alkaline phosphatase activity and related enzymes in diabetes was reported (Goldberg et al, 1977; Maxwell et al, 1986) supporting an association between the severity of

			diabetes and diabetic bone disease.
114	**Aspartate aminotransferase**	**Serum aspartate aminotransferase AST (IU/L)**	**Al-Jameil et al (2014) found that elevated ALT and GGT were significantly, whereas elevated AST was only marginally, positively associated with T2DM.**
115	**Alanine aminotransferase**	**Serum alanine aminotransferase ALT (IU/L)**	**Fei et al (2012) reported that elevated serum ALT and GGT were positively associated with prediabetes and insulin resistance. Nguyen et al (2011) found that elevated ALT and GGT levels could be potential biomarkers of risk of incident prediabetes and diabetes. Ko et al (2015) found that elevated GGT and ALT and lower AST/ALT within the physiological range were independent, additive risk factors of T2DM and IFG; a precursor stage of overt diabetes.** **Kubo et al (2007) reported that serum GGT and ALT concentrations are strong predictors of diabetes in the general population, independent of known risk factors. Yokota et al (2017) found in a predictive modelling study that elevated ALT was a significant, independent predictor of conversion from prediabetes to diabetes.**
116	Bicarbonate	Plasma bicarbonate (mmol/L)	Electrolyte and acid–base disturbances are observed in diabetic patients (Palmer & Clegg, 2015). A study reported that low urine pH is an independent predictor of diabetes and can be an easy practical marker for diabetes (Hashimoto et al, 2017).
117	**Calcium**	**Serum total calcium (mg/dL)**	**Rooney et al (2016) found that elevated serum calcium was associated with a greater risk of T2DM. Suh et al (2017) revealed that the elevation of albumin-adjusted serum calcium levels was associated with an increased risk of T2DM, independent of baseline glycemic status. A meta-analysis conducted by Sing et al (2016) found that elevated serum calcium concentration is associated with incident diabetes. Fu et al (2015) revealed that the prevalence of diabetes increased significantly with serum calcium level. However, Zaccardi et al (2015) observed no such association between direct measurement of active calcium and risk of T2DM.**
118	Creatine phosphokinase	Serum creatine phosphokinase(CPK) (IU/L)	Elevated serum creatine phosphokinase levels are observed in patients with skeletal muscle infarction induced by diabetes (Grigoriadis et al, 2000). Compartmentation of hexokinase and creatine phosphokinase is regulated by insulin (Bessman & Geiger, 1980).
119	Chloride	Blood chloride (mmol/L)	Electrolyte and acid–base disturbances are observed in diabetic patients (Palmer & Clegg, 2015).
120	Creatinine	Serum creatinine (mg/dL)	Glomerular hyperfiltration was reported in prediabetes and prehypertension in a study which used serum creatinine to determine the glomerular filtration rate (Okada et al, 2011). Another study found that uurinary albumin: creatinine ratio

			predicts prediabetes progression to diabetes and reversal to normoglycemia (Dutta et al, 2014).
121	Globulin	Serum globulin (g/dL)	A study found that low sex hormone-binding globulin levels are associated with prediabetes in Chinese men independent of total testosterone (Zhu et al, 2016). In another study decreased IgG and IgM, and increased IgA levels were independently associated with the prevalence of T2DM among the adult population indicating that the immunoglobulins might be useful predictive factors for T2DM in the general adult population (Guo et al, 2016). An association between immunoglobulin concentrations and prediabetes prevalence was reported in a large Chinese cohort (Wang, H. et al, 2017).
122	**Gamma glutamyl transferase**	**Serum gamma glutamyl transferase (U/L)**	**See also ALT. A Mendelian randomization study by Lee et al (2016) found weak genetic evidence that GGT levels may have a causal role in the development of T2DM. Another Mendelian randomization study by Conen et al (2010) however found evidence for a direct causal relation of GGT with fasting insulin. André et al (2007) found GGT, a predictor of type 2 diabetes, was associated with a risk of incident metabolic syndrome and that the association was mainly related with insulin resistance but was independent of other confounding factors.**
123	Iron	Iron, refrigerated serum (ug/dL)	Mildly elevated body iron stores are associated with statistically significant elevations in glucose homeostasis indexes (Tuomainen et al, 1997). Serum ferritin was found to be a biomarker of clinically incident diabetes in a study done by Salomaa et al (2010).
124	**Potassium**	**Serum potassium (mmol/L)**	**Meisinger et al (2013) revealed that serum potassium levels were independently associated with prediabetes. Chatterjee et al (2010) found that low serum potassium level is an independent predictor of incident diabetes. Chatterjee et al (2011) reviewed literature on the association between the low serum and dietary potassium levels and incident diabetes. Heianza et al (2011) revealed mild to moderately low serum potassium levels, within the normal range and without frank hypokalaemia, could be predictive of T2DM in apparently healthy Japanese men.**
125	Lactate dehydrogenase	Serum lactate dehydrogenase (U/L)	An experimental study revealed that acute overexpression of lactate dehydrogenase-A perturbs cell mitochondrial metabolism and insulin secretion (Ainscow et al, 2000). Elevated salivary lactate dehydrogenase levels were detected in diabetic patients (Musumeci et al, 1993).
126	Sodium	Plasma sodium (mmol/L)	Electrolyte and acid–base disturbances are observed in diabetic patients (Palmer & Clegg, 2015). Inverse distributions of serum sodium and potassium were observed in uncontrolled diabetic in-patients (Saito et al, 1999). Diabetes is associated with both hypo-

			and hyper-natremia demonstrating the coexistence of varying hyperglycemia-related mechanisms (Liamis et al, 2014).
127	**Osmolality**	**Osmolality (mmol/Kg)**	**Salazar Vázquez et al (2010) found that normal children compensate for the increase in vascular resistance due to increased blood viscosity (increased hemoglobin and hematocrit) while diabetic children do not, probably due to endothelial dysfunction indicating a direct correlation between blood pressure and blood viscosity in diabetes type 1 children but not in normal. Irace et al (2014) found a direct relationship between blood viscosity and blood glucose in nondiabetic subjects and that even within glucose values considered completely normal, individuals with higher blood glucose levels have increased blood viscosity comparable with that observed in subjects with prediabetes. Marini et al (2017) showed that individuals with HbA1c-defined prediabetes have increased predicted blood viscosity.**
128	Phosphorus	Serum phosphorus (mg/dL)	Disturbance of inorganic phosphate metabolism are observed in diabetes and clinical manifestations of phosphorus-depletion syndrome may be seen during recovery from diabetic ketoacidosis (Ditzel et al, 2010). An inverse association was reported between serum phosphorus levels and diabetes in adults (Fang & Li, 2016)
129	Bilirubin	Total bilirubin (mg/dL)	Serum bilirubin is positively associated with incident diabetes (Lee et al, 2017)
130	Protein	Total protein (g/dL)	A matched case-control study reported that total protein levels were statistically significantly increased compared to controls in diabetic patients (Venkataramana et al, 2013). Increased levels of acute-phase serum proteins are observed in diabetes (McMillan, 1989). The assay for glucosylated serum protein is a sensitive indicator of the degree of hyperglycemia in diabetes (McFarland et al, 1979).
131	**Triglyceride**	**Triglycerides, refrigerated (mg/dL)**	**Deepa et al (2015) found that high cholesterol, high triglyceride, and low HDL cholesterol levels were each independently associated with both prediabetes and diabetes. Al-Aubaidy & Jelinek (2014) found oxidative stress and triglycerides were predictors of subclinical atherosclerosis in prediabetes. Gao et al (2017) found that the aggravation of serum triglyceride level was related to diabetic progression. Abbasi et al (2016) found that fasting TG concentration in individuals with prediabetes is a marker of high risk for insulin resistance, CHD, and diabetes. Akehi et al (2010) found that elevated levels of GGT, TC and TG are good clinical markers to predict diabetic risks, even in young NGT males. Calanna et al (2014) found that subjects with HbA1c-defined prediabetes and type 2 diabetes, respectively, are characterized by abnormalities**

			in lipid profile and liver steatosis, thus exhibiting a severe risk profile for cardiovascular and liver diseases. Hilawe et al (2016) revealed that elevated TG level is a significant predictor of diabetes. Okwechime et al, 2015 found significant positive associations between hypercholesterolemia and prediabetes and diabetes both.
132	**Serum uric acid (SUA)**	**Uric acid (mg/dL)**	**Anothaisintawee et al (2017) found that an increase in SUA was directly and indirectly associated with increased FPG but the effect of SUA on HbA1c was shown when it was mediated through waist circumference (WC). Vučak et al (2012) found a positive association between hyperuricemia and prediabetes. Chu et al (2017) found that there was a remarkable association of uric acid calculi with prediabetic and diabetic states. Zhang, Q. et al (2016) revealed that mean SUA value was strongly and positively related to prediabetes risk and showed better predictive ability for prediabetes than baseline SUA. Jia et al (2013) reported that SUA levels are positively associated with incidence of IFG and T2DM, and the association might be nonlinear. Van Der Schaft et al (2017) reported findings that were consistent with the notion that serum uric acid is more closely related to early-phase mechanisms in the development of type 2 diabetes mellitus than late-phase mechanisms.**
133	Tissue transglutaminase	Tissue transglutaminase (IgA TTG); Dichotomized as negative versus positive/weakly positive	An association between prevalence of tissue transglutaminase IgA antibodies and T1DM is reported (Gabriel et al, 2011).
134	Vitamin B12 level	Vitamin B12(pg/mL)	The B group vitamins, thiamin, pyridoxine, and biotin have been found decreased in T2DM patients, but the mechanism is inconclusive (Valdés-Ramos et al, 2015). Amer et al (2015) investigated link between vitamin B12, T2DM, and bone mineral density in elderly patients
Newly created/modified/composite variables			
135	Education	Codes were ascribed as follows: 1 = < 9^{th} grade, 2 = 9-11 grade, 3 = high school, 4 = college/ AA degree/ college graduate/ above	A study on the impact of socio-economic factors on diabetes found a significant inverse relationship between maternal education and diabetes in late life of adult offspring. Individuals with better educated mothers have a lower risk of being diabetic after age 50. This relationship remains after controlling for other childhood and adult risk factors (Kohler & Soldo, 2005). Socio-economic factors are strong predictors of health-related quality of life among diabetic patients (Shamshirgaran et al, 2014). Lower educational level is a predictor of incident type 2 diabetes in European countries (Sacerdote et al, 2012)
136	Processed food expenditure	Money spent on junk food as a ratio of total expenditure per	Among offsprings of parents with T2DM, self-reported dietary habits correlate with measures of

		month. Calculated as follows: (CBD120 + CBD 130)/ (CBD070 + CBD090 + CBD110 + CBD120 + CBD130) Variables: CBD120 = Money spent on eating out past 30 days, CBD130 = Money spent on carryout/delivered foods past 30 days, CBD070 = Money spent at supermarket/grocery store past 30 days, CBD090 = Money spent on nonfood items past 30 days, CBD110 = Money spent on food at other stores past 30 days	adiposity and dyslipidemia (Boucher et al, 2015). Rising consumption of refined carbohydrate and decreasing intakes of fiber paralleled the upward trend in the prevalence of T2DM in the US during the 20th century (Gross et al, 2004). A meta-analysis revealed that processed meat intake is a dietary risk factor for diabetes (Micha et al, 2010).
137	Creatine kinase	A dichotomous variable was created as follows: 1 = no: If responded "no" to all the 4 questions below 2 = yes: If responded "yes" to any one or more or all the 4 questions below; Last 3 days strenuous exercise = yes OR exercise made muscles sore/painful = yes OR injury made muscles sore/painful = yes OR other muscle pain/soreness last 3 days = yes	Creatine kinase, a biomarker of muscular injury, is positively associated with diabetes (Jevrić-Čaušević et al, 2006; Lazarov et al, 1990). Bartlett et al (2017) revealed an association of the composite inflammatory biomarker GlycA, with exercise-induced changes in body habitus of prediabetic individuals.
138	Number of processed food meals	Sum of the number of meals not home prepared (i.e. prepared in places such as restaurants, fast food places, food stands or from vending machines), the number of ready to eat foods, and the number of frozen meals/pizza during the past 30 days Calculated by summing up the following variables: (DBD895*4) + DBD905 + DBD910 DBD895 = # of meals not home prepared last 7 days * 4, DBD905 = # of ready-to-eat foods in past 30 days, DBD910 = # of frozen meals/pizza in past 30 days NB: Since DBD895 gives the number of meals not home prepared last 7 days, the figure was multiplied by 4 to estimate the monthly figure.	See no:136 above

139	Disability	A new variable was created as follows: 1 = no; If responded "no" to all the 6 questions below 2 = yes: If responded "yes" to any one or more or all the 6 questions; Have serious difficulty hearing, Have serious difficulty seeing, Have serious difficulty concentrating, Have serious difficulty walking, Have difficulty dressing or bathing, and Have difficulty doing errands alone Corresponding NHANES variables: DLQ010 = Do you have serious difficulty hearing? DLQ020 = Do you have serious difficulty seeing? DLQ040 = Do you have serious difficulty concentrating? DLQ050 = Do you have serious difficulty walking? DLQ060 = Do you have difficulty dressing or bathing? DLQ080 = Do you have difficulty doing errands alone?	Functional disability and depressive symptoms were strongly associated with diabetes (Kalyani et al, 2017)
140	HPV vaccination	NHANES database provides HPV vaccination history for male and female participants separately. A new dichotomous variable (1 = no, 2 = yes) was created by merging the responses to the two variables IMQ040 (females' self-reported HPV vaccination history) and IMQ070 (males' self-reported HPV vaccination history)	A positive association between genital warts and diabetes has been reported (Yong et al, 2010). Also, an association between diabetes/hyperglycemia and the prognosis of cervical cancer patients was found by a meta-analysis (Chen et al, 2017).
141	Urine leakage	A dichotomous variable was created as follows: 1 = no; If responded "no" to all the 3 questions below 2 = yes: If responded yes to any one or more or all the 3 questions; Leak urine during physical activities, Urinated before reaching the toilet, Leak urine during nonphysical activities Corresponding NHANES variables: KIQ042 = During the past 12 months, have you leaked urine during physical	See no:42

		activity? KIQ044 = During the past 12 months, have you urinated before you could reach the toilet? KIQ046 = During the past 12 months, have you leaked urine during non-physical activities?	
142	Depression	The NHANES uses the Patient Health Questionnaire (Kroenke & Spitzer, 2002; Kroenke et al, 2001), a 9-item depression screening tool, to determine the prevalence of depression. Response categories for the nine-item instrument "not at all," "several days," "more than half the days," and "nearly every day" are given a score ranging from 0 to 3. The DSM-IV depression diagnostic criteria (Spitzer et al.1999) can be used along with the tool to categorize into groups as follows: Depression severity: 0-4 none, 5-9 mild, 10-14 moderate, 15-19 moderately severe, 20-27 severe. However, in the present study, the depression screening score was included as a continuous variable.	Comorbid depression was found to be a predictor of health-related quality of life among people with T2DM (Shamshirgaran et al, 2014). Another study revealed that depression and anxiety were simultaneous predictors of diabetes incidence (Khambaty et al, 2017). Also, comorbid depression and functional disability were associated with diabetes (Kalyani et al, 2017)
143	Pesticide	A dichotomous variable was created as follows: 1 = no: If responded "no" to both questions below 2 = yes: If responded "yes" to either of the 2 questions; products used in home to control insects, products used to kill weeds Corresponding NHANES variables are: PUQ100 = Were any chemical products used in your home to control fleas, roaches, ants, termites, or other insects past 7 days? PUQ110 = Were any chemical products used in your lawn or garden to kill weeds past 7 days?	Pesticide use, especially organochlorine and organophosphate insecticides, were positively associated with incident diabetes in a prospective cohort study (Montgomery et al, 2008). Associations between biomarkers of pesticide exposure and diabetes were revealed in a study based on the NHANES 1999–2004 data (Everett & Matheson, 2010).
144	**Vigorous activity**	**A dichotomous variable was created as follows:** **1 = no: If responded "no" to both questions below** **2 = yes: If responded "yes"**	**Jung et al (2015) found that high-intensity interval training is an efficacious alternative to moderate-intensity continuous training for adults with prediabetes. Malin et al, 2012 revealed independent and combined effects of exercise**

		to any one or both the questions pertaining to; vigorous work activity, vigorous recreational activities **Corresponding NHANES variables: PAQ605 = Does your work involve vigorous-intensity activity that causes large increases in breathing or heart rate like carrying or lifting heavy loads, digging or construction work for at least 10 minutes continuously? PAQ650 = In a typical week do you do any vigorous-intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously?**	**training and metformin on insulin sensitivity in prediabetic individuals. Earnest (2008) proposed exercise interval training as an improved stimulus for improving the physiology of prediabetes. Liu et al (2013) found that exercise interventions may reverse vascular endothelium-dependent dysfunction in middle-aged patients with impaired glucose tolerance providing direct clinical evidence supporting the use of exercise intervention to prevent diabetes during the early stage. Anjana et al (2015) found that physical inactivity is a predictor of progression to dysglycemia in a population-based Asian-Indian cohort. Boucher et al (2015) found that among African-American and Caucasian offspring of parents with T2DM, self-reported dietary and exercise habits correlated with measures of adiposity and dyslipidemia; however, physical activity, but not dietary recall, significantly predicted incident dysglycemia during 5.5 years of follow-up.**
145	Moderate activity	A dichotomous variable was created as follows: 1 = no: If responded "no" to both questions below 2 = yes: If responded "yes" to any one or both the questions pertaining to; moderate work activity, moderate recreational activities Corresponding NHANES variables: PAQ620 = Does your work involve moderate-intensity activity that causes small increases in breathing or heart rate such as brisk walking or carrying light loads for at least 10 minutes continuously? PAQ665 = In a typical week do you do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously?	Physical inactivity is a predictor of progression to dysglycemia (Anjana et al, 2015). Exercise habits correlate with measures of adiposity and dyslipidemia and physical inactivity is a predictor of incident dysglycemia (Boucher et al, 2015).
146	Functional limitation	A dichotomous variable was created as follows: 1 = no; If responded "no" to all the questions below 2 = yes: If responded "yes" to any of the questions pertaining	See no:142, no:144, and no:145 above. Frailty is an independent risk factor for incident T2DM in the elderly (Veronese et al, 2016).

		to: Crawl, walk, run, play limitations; Impairment requiring special equipment; Limitations keeping you from working; Limited in amount of work you can do; Need special equipment to walk; Experience confusion/memory problems; Physical, mental, emotional limitations Corresponding NHANES variables: PFQ020 = Do you have an impairment or health problem that limits your ability to walk, run or play? PFQ033 = Do you have any impairment or health problem that requires you to use special equipment, such as a brace, a wheelchair, or a hearing aid? PFQ049 = Does a physical, mental or emotional problem now keep you from working at a job or business? PFQ051 = Are you/ limited in the kind or amount of work you can do because of a physical, mental, or emotional problem? PFQ054 = Because of a health problem, do you have difficulty walking without using any special equipment? PFQ057 = Are you limited in any way because of difficulty remembering or because you experience periods of confusion? PFQ059 = Are you limited in any way in any activity because of a physical, mental, or emotional problem?	
147	Sleeping trouble	A dichotomous variable was created as follows: 1 = no: If responded "no" to both SLQ050 and SLQ060 2 = yes: If responded "yes" to either SLQ050 or SLQ060 Relevant NHANES variables: SLQ050 = Have you ever told a doctor or other health professional that you have trouble sleeping? SLQ060 = Have you ever been told by a doctor or other	See no:77 above

		health professional that you have a sleep disorder?	
148	Secondhand smoking	A dichotomous variable was created as follows: 1 = no: If responded "no" to all questions below 2 = yes: If responded "yes" to one or more or all 6 questions; SMQ858, SMQ862, SMQ868, SMQ872, SMQ876, SMQ880 Relevant NHANES variables: SMQ858 = Last 7 days, while you were working at a job or business outside of the home, did someone else smoke cigarettes or other tobacco products indoors? SMQ862 = Last 7 days, while you were in a restaurant, did someone else smoke cigarettes or other tobacco products indoors? SMQ868 = Last 7 days, while you were in a bar, did someone else smoke cigarettes or other tobacco products indoors? SMQ872 = Last 7 days, while you were riding in a car or motor vehicle, did someone else smoke cigarettes or other tobacco products? SMQ876 = Last 7 days, while you were in a home other than your own, did someone else smoke cigarettes or other tobacco products indoors? SMQ880 = Last 7 days, while you were in the other indoor area, did someone else smoke cigarettes or other tobacco products?	See no:78, no:79 and no:80 above. Both firsthand and secondhand smoking are risk factors of T2DM. The risk of diabetes is elevated in new quitters but decreases substantially as the time since quitting increases (Pan et al, 2015).
149	Gestational DM	NHANES question RHQ162: During your pregnancy, were you ever told by a doctor or other health professional that you had diabetes, sugar diabetes or gestational diabetes? Responses were recoded as 1 = no, 2 = borderline/yes Since the responses for this question were collected only from females 20 years of age and above, the multiple	A question on gestational DM is included in the ADA prediabetes screening tool as it is a well-established risk factor for prediabetes (ADA diabetes screening test, 2018). Gestational DM was found to be a risk factor for subsequent development of prediabetes/diabetes and a hyperglycemic intrauterine environment seemed to be involved in the pathogenesis (Clausen et al, 2008).

		imputation was performed for missing data among that target group. Other missing data were categorized as "no"	
150	Overweight baby at birth	NHANES question RHQ172: Did your delivery/Did any of your deliveries result in a baby that weighed 9 pounds (4082 g) or more at birth? Responses were recoded as 1 = no, 2 = yes Since the responses for this question were collected only from females 20 years of age and above, the multiple imputation was performed for missing data among that target group. Other missing data were categorized as "no"	A question on having an overweight baby at birth is included in the CDC prediabetes screening test as it is a strong predictor of prediabetes (CDC prediabetes screening test, 2018).
151	**Hysterectomy**	**NHANES question RHD280: Have you had a hysterectomy that is, surgery to remove your uterus or womb?** **Responses were recoded as 1 = no, 2 = yes** **Since the responses for this question were collected only from females 20 years of age and above, the multiple imputation was performed for missing data among that target group. Other missing data were categorized as "no"**	**Luo et al (2017) observed that hysterectomy, regardless of oophorectomy status, was associated with increased risk of diabetes among postmenopausal women. Wilson & Mishra (2017) found an association between hysterectomy and the incidence of diabetes mediated through BMI. Appiah et al (2014) reported that women with hysterectomy concomitant with bilateral oophorectomy (BSO) status may represent a unique population with elevated risk for diabetes.**
152	Bilateral ovariectomy	NHANES question RHQ305: Have you had both of your ovaries removed (either when you had your uterus removed or at another time)? Responses were recoded as 1 = no, 2 = yes Since the responses for this question were collected only from females 20 years of age and above, the multiple imputation was performed for missing data among that target group. Other missing data were categorized as "no"	Bilateral oophorectomy increases the risk of incident diabetes in postmenopausal women (Appiah et al, 2014). Similar and additive effects of ovariectomy and diabetes on insulin resistance and lipid metabolism have been revealed (Tawfik et al, 2015).
153	Oral contraception	NHANES question RHQ420: Have you ever taken birth control pills for any reason? Responses were recoded as 1 = no, 2 = yes Since the responses for this	In a hospital-based study among diabetic patients, steroid use showed a steady threefold increase in odds for suboptimal glycemic control across all rates of use (Bender et al, 2015). Hormonal contraceptive intake has been shown to adversely affect glycemic homeostasis (Cortés & Alfaro,

		question were collected only from females 12 years of age and older, the multiple imputation was performed for missing data among that target group. Other missing data were categorized as "no"	2014). The effects of different combined oral contraceptives on glucose tolerance test glucose and insulin profiles may be due to a combination of estrogen-induced insulin resistance and progestin-associated changes in insulin half-life (Godsland et al, 1992)
154	Female hormones intake	NHANES question RHQ540: "Have you ever used female hormones such as estrogen and progesterone? Please include any forms of female hormones, such as pills, cream, patch, and injectables, but do not include birth control methods or use for infertility." Responses were recoded as 1 = no, 2 = yes Since the responses for this question were collected only from females 20 years of age and older, the multiple imputation was performed for missing data among that target group. Other missing data were categorized as "no"	See no:153 above. Hormone replacement therapy was associated with increased insulin sensitivity and may improve insulin resistance and glucose homeostasis in women with diabetes (Bitoska et al, 2016).
155	**Mean SBP**	**Average value of the 3 or 4 SBP readings made on each participant of the study (BPXSY1, BPXSY2, BPXSY3, BPXSY4)**	**Hypertension is an independent predictor of prediabetes (Casapulla et al, 2017; Satman et al, 2013; Anjana et al, 2011; Soewondo & Pramono, 2011)**
156	Mean DBP	Average value of the 3 or 4 DBP readings made on each participant of the study (BPXDI1, BPXDI2, BPXDI3, BPXDI4)	See no: 155 above

References

Abbasi, F., Kohli, P., Reaven, G.M. and Knowles, J.W., 2016. Hypertriglyceridemia: A simple approach to identify insulin resistance and enhanced cardio-metabolic risk in patients with prediabetes. *diabetes research and clinical practice*, *120*, pp.156-161.

ADA diabetes screening test, 2018. Available online at http://main.diabetes.org/dorg/PDFs/risk-test-paper-version.pdf

Aeschbacher, S., Schoen, T., Clair, C., Schillinger, P., Schönenberger, S., Risch, M., Risch, L. and Conen, D., 2014. Association of smoking and nicotine dependence with pre-diabetes in young and healthy adults. *Swiss Med Wkly*, *144*, p.w14019.

Ainscow, E.K., Zhao, C. and Rutter, G.A., 2000. Acute overexpression of lactate dehydrogenase-A perturbs beta-cell mitochondrial metabolism and insulin secretion. *Diabetes*, *49*(7), pp.1149-1155.

Akehi, Y., Tsutsumi, Y., Tatsumoto, A., Yoshida, R., Ohkubo, K., Takenoshita, H., Kudo, T., Ashida, K., Anzai, K., Yamashita, T. and Kawashima, H., 2010. Serum γ-glutamyltransferase, triglyceride and total cholesterol are possible prediabetic risk markers in young Japanese men. *Endocrine journal*, *57*(11), pp.981-989.

Al Shehri, Z.S., 2017. The relationship between some biochemical and hematological changes in type 2 diabetes mellitus. *Biomedical Research and Therapy*, *4*(11), pp.1760-1774.

Al-Aubaidy, H.A. and Jelinek, H.F., 2014. Oxidative stress and triglycerides as predictors of subclinical atherosclerosis in prediabetes. *Redox Report*, *19*(2), pp.87-91.

Ali, A.A.E.S., El Deeb, G.S., Essa, A.A.S. and Ahmed, N.S.S., 2014. Study of frequency of prediabetes in Egyptian patients with chronic hepatitis C virus infection. *Menoufia Medical Journal*, *27*(2), p.453.

Al-Jameil, N., Khan, F.A., Arjumand, S., Khan, M.F. and Tabassum, H., 2014. Associated liver enzymes with hyperlipidemic profile in type 2 diabetes patients. *International journal of clinical and experimental pathology*, *7*(7), p.4345.

Amer, M.S., Ali-Labib, R., Farid, T.M., Rasheedy, D. and Tolba, M.F., 2015. Link between vitamin B12, type 2 diabetes mellitus, and bone mineral density in elderly patients. *Journal of Clinical Gerontology and Geriatrics*, *6*(4), pp.120-124.

Anderson, R.J., Freedland, K.E., Clouse, R.E. and Lustman, P.J., 2001. The prevalence of comorbid depression in adults with diabetes: a meta-analysis. *Diabetes care*, *24*(6), pp.1069-1078.

André, P., Balkau, B., Charles, M.A. and Eschwège, E., 2007. γ-glutamyltransferase activity and development of the metabolic syndrome (International Diabetes Federation Definition) in middle-aged men and women: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR) cohort. *Diabetes Care*, *30*(9), pp.2355-2361.

Anjana, R.M., Pradeepa, R., Deepa, M., Datta, M., Sudha, V., Unnikrishnan, R., Bhansali, A., Joshi, S.R., Joshi, P.P., Yajnik, C.S. and Dhandhania, V.K., 2011. Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research–INdia DIABetes (ICMR–INDIAB) study. *Diabetologia*, *54*(12), pp.3022-3027.

Anjana, R.M., Rani, C.S.S., Deepa, M., Pradeepa, R., Sudha, V., Nair, H.D., Lakshmipriya, N., Subhashini, S., Binu, V.S., Unnikrishnan, R. and Mohan, V., 2015. Incidence of diabetes and prediabetes and predictors of progression among Asian Indians: 10-year follow-up of the Chennai Urban Rural Epidemiology Study (CURES). *Diabetes Care*, p.dc142814.

Anothaisintawee, T., Lertrattananon, D., Thamakaison, S., Reutrakul, S., Ongphiphadhanakul, B. and Thakkinstian, A., 2017. Direct and Indirect Effects of Serum Uric Acid on Blood Sugar Levels in Patients with Prediabetes: A Mediation Analysis. *Journal of diabetes research*, *2017*.

Anothaisintawee, T., Reutrakul, S., Van Cauter, E. and Thakkinstian, A., 2016. Sleep disturbances compared to traditional risk factors for diabetes development: systematic review and meta-analysis. *Sleep medicine reviews*, *30*, pp.11-24.

Appiah, D., Winters, S.J. and Hornung, C.A., 2014. Bilateral oophorectomy and the risk of incident diabetes in postmenopausal women. *Diabetes care*, *37*(3), pp.725-733.

Armstrong, A.W., Harskamp, C.T. and Armstrong, E.J., 2013. Psoriasis and the risk of diabetes mellitus: a systematic review and meta-analysis. *JAMA dermatology*, *149*(1), pp.84-91.

Bartlett, D.B., Slentz, C.A., Connelly, M.A., Piner, L.W., Willis, L.H., Bateman, L.A., Granville, E.O., Bales, C.W., Huffman, K.M. and Kraus, W.E., 2017. Association of the Composite Inflammatory Biomarker GlycA, with Exercise-Induced Changes in Body Habitus in Men and Women with Prediabetes. *Oxidative medicine and cellular longevity*, *2017*.

Bassareo, P.P., Marras, A.R., Barbanti, C. and Mercuro, G., 2013. Comparison between waist and mid-upper arm circumferences in influencing systolic blood pressure in adolescence: the SHARP (Sardinian Hypertensive Adolescent Research Programme) study. *Journal of Pediatric and Neonatal Individualized Medicine (JPNIM)*, *2*(2), p.e020207.

Bender, M., Smith, T.C., Thompson, J., Koucheki, A. and Holdy, K., 2015. Predictors of suboptimal glycemic control for hospitalized patients with diabetes: Targets for clinical action.

Bessman, S.P. and Geiger, P.J., 1980. Compartmentation of hexokinase and creatine phosphokinase, cellular regulation, and insulin action. In *Current topics in cellular regulation* (Vol. 16, pp. 55-86). Academic Press.

Binh, T.Q., Thu, N.T.T., Phuong, P.T., Nhung, B.T. and Nhung, T.T.H., 2015. CDKN2A-rs10811661 polymorphism, waist-hip ratio, systolic blood pressure, and dyslipidemia are the independent risk factors for prediabetes in a Vietnamese population. *BMC genetics*, *16*(1), p.107.

Bitoska, I., Krstevska, B., Milenkovic, T., Subeska-Stratrova, S., Petrovski, G., Mishevska, S.J., Ahmeti, I. and Todorova, B., 2016. Effects of hormone replacement therapy on insulin resistance in postmenopausal diabetic women. *Open access Macedonian journal of medical sciences*, *4*(1), p.83.

Boucher, A.B., Adesanya, E.O., Owei, I., Gilles, A.K., Ebenibo, S., Wan, J., Edeoga, C. and Dagogo-Jack, S., 2015. Dietary habits and leisure-time physical activity in relation to adiposity, dyslipidemia, and incident dysglycemia in the pathobiology of prediabetes in a biracial cohort study. *Metabolism-Clinical and Experimental*, *64*(9), pp.1060-1067.

Bulum, T., Kolarić, B. and Duvnjak, L., 2014. Decreased serum monocytes and elevated neutrophils as additional markers of insulin resistance in type 1 diabetes. *International Journal of Diabetes in Developing Countries*, *34*(3), pp.150-155.

Burman, B.E., Bacchetti, P., Ayala, C.E., Gelman, N., Melgar, J. and Khalili, M., 2015. Liver inflammation is a risk factor for prediabetes in at-risk latinos with and without hepatitis C infection. *Liver International*, *35*(1), pp.101-107.

Calanna, S., Scicali, R., Di Pino, A., Knop, F.K., Piro, S., Rabuazzo, A.M. and Purrello, F., 2014. Lipid and liver abnormalities in haemoglobin A1c-defined prediabetes and type 2 diabetes. *Nutrition, Metabolism and Cardiovascular Diseases*, *24*(6), pp.670-676.

Casapulla, S.L., Howe, C.A., Mora, G.R., Berryman, D., Grijalva, M.J., Rojas, E.W., Nakazawa, M. and Shubrook, J.H., 2017. Cardiometabolic risk factors, metabolic syndrome and pre-diabetes in adolescents in the Sierra region of Ecuador. *Diabetology & metabolic syndrome*, *9*(1), p.24.

CDC prediabetes screening test, 2018. Available online at https://www.cdc.gov/diabetes/prevention/pdf/prediabetestest.pdf

Charfen, M.A., Ipp, E., Kaji, A.H., Saleh, T., Qazi, M.F. and Lewis, R.J., 2009. Detection of undiagnosed diabetes and prediabetic states in high-risk emergency department patients. *Academic Emergency Medicine*, *16*(5), pp.394-402.

Chatterjee, R., Yeh, H.C., Edelman, D. and Brancati, F., 2011. Potassium and risk of type 2 diabetes. *Expert review of endocrinology & metabolism*, *6*(5), pp.665-672.

Chatterjee, R., Yeh, H.C., Shafi, T., Selvin, E., Anderson, C., Pankow, J.S., Miller, E. and Brancati, F., 2010. Serum and dietary potassium and risk of incident type 2 diabetes mellitus: The Atherosclerosis Risk in Communities (ARIC) study. *Archives of internal medicine*, *170*(19), pp.1745-1751.

Chen, S., Tao, M., Zhao, L. and Zhang, X., 2017. The association between diabetes/hyperglycemia and the prognosis of cervical cancer patients: A systematic review and meta-analysis. *Medicine*, *96*(40).

Chesla, C.A., Fisher, L., Skaff, M.M., Mullan, J.T., Gilliss, C.L. and Kanter, R., 2003. Family predictors of disease management over one year in Latino and European American patients with type 2 diabetes. *Family Process*, *42*(3), pp.375-390.

Choi, H.K., De Vera, M.A. and Krishnan, E., 2008. Gout and the risk of type 2 diabetes among men with a high cardiovascular risk profile. *Rheumatology*, *47*(10), pp.1567-1570.

Chu, F.Y., Chang, C.C., Huang, P.H., Lin, Y.N., Ku, P.W., Sun, J.T., Ho, J.L., Yen, T.H. and Su, M.J., 2017. The Association of Uric Acid Calculi with Obesity, Prediabetes, Type 2 Diabetes Mellitus, and Hypertension. *BioMed research international*, *2017*.

Clausen, T.D., Mathiesen, E.R., Hansen, T., Pedersen, O., Jensen, D.M., Lauenborg, J. and Damm, P., 2008. High prevalence of type 2 diabetes and pre-diabetes in adult offspring of women with gestational diabetes mellitus or type 1 diabetes: the role of intrauterine hyperglycemia. *Diabetes care*, *31*(2), pp.340-346.

Conen, D., Vollenweider, P., Rousson, V., Marques-Vidal, P., Paccaud, F., Waeber, G. and Bochud, M., 2010. Use of a Mendelian randomization approach to assess the causal relation of γ-glutamyltransferase with blood pressure and serum insulin levels. *American journal of epidemiology*, *172*(12), pp.1431-1441.

Cortés, M.E. and Alfaro, A.A., 2014. The effects of hormonal contraceptives on glycemic regulation. *The Linacre Quarterly*, *81*(3), pp.209-218.

Cruickshank, K., Riste, L., Anderson, S.G., Wright, J.S., Dunn, G. and Gosling, R.G., 2002. Aortic pulse-wave velocity and its relationship to mortality in diabetes and glucose intolerance: an integrated index of vascular function?. *Circulation*, *106*(16), pp.2085-2090.

Cullmann, M., Hilding, A. and Östenson, C.G., 2012. Alcohol consumption and risk of pre-diabetes and type 2 diabetes development in a Swedish population. *Diabetic Medicine*, *29*(4), pp.441-452.

Dasbach, E.J., Klein, R., Klein, B.E. and Moss, S.E., 1994. Self-rated health and mortality in people with diabetes. *American Journal of Public Health*, *84*(11), pp.1775-1779.

Davis, J.A., Tsui, I., Gelberg, L., Gabrielian, S., Lee, M.L. and Chang, E.T., 2017. Risk factors for diabetic retinopathy among homeless veterans. *Psychological services*, *14*(2), p.221.

Davis, N.J., Ma, Y., Delahanty, L.M., Hoffman, H.J., Mayer-Davis, E., Franks, P.W., Brown-Friday, J., Isonaga, M., Kriska, A.M., Venditti, E.M. and Wylie-Rosett, J., 2013. Predictors of sustained reduction in energy and fat intake in the Diabetes Prevention Program Outcomes Study intensive lifestyle intervention. *Journal of the Academy of Nutrition and Dietetics*, *113*(11), pp.1455-1464.

Deepa, M., Grace, M., Binukumar, B., Pradeepa, R., Roopa, S., Khan, H.M., Fatmi, Z., Kadir, M.M., Naeem, I., Ajay, V.S. and Anjana, R.M., 2015. High burden of prediabetes and diabetes in three large cities in South Asia: The Center for cArdio-metabolic Risk Reduction in South Asia (CARRS) Study. *Diabetes research and clinical practice*, *110*(2), pp.172-182.

Devereux, R.B., Roman, M.J., Paranicas, M., O'grady, M.J., Lee, E.T., Welty, T.K., Fabsitz, R.R., Robbins, D., Rhoades, E.R. and Howard, B.V., 2000. Impact of diabetes on cardiac structure and function: the strong heart study. *Circulation*, *101*(19), pp.2271-2276.

Ditzel, J. and Lervang, H.H., 2010. Disturbance of inorganic phosphate metabolism in diabetes mellitus: clinical manifestations of phosphorus-depletion syndrome during recovery from diabetic ketoacidosis. *Diabetes, metabolic syndrome and obesity: targets and therapy*, *3*, p.319.

Donahue, R.P., Rejman, K., Rafalson, L.B., Dmochowski, J., Stranges, S. and Trevisan, M., 2007. Sex differences in endothelial function markers before conversion to pre-diabetes: Does the clock start ticking earlier among women?: The western New York study. *Diabetes Care*, *30*(2), pp.354-359.

Dutta, D., Choudhuri, S., Mondal, S.A., Mukherjee, S. and Chowdhury, S., 2014. Urinary albumin: creatinine ratio predicts prediabetes progression to diabetes and reversal to normoglycemia: Role of associated insulin resistance, inflammatory cytokines and low vitamin D. *Journal of diabetes*, *6*(4), pp.316-322.

Earnest, C.P., 2008. Exercise interval training: an improved stimulus for improving the physiology of pre-diabetes. *Medical hypotheses*, *71*(5), pp.752-761.

Elwood, P.C., Pickering, J.E. and Fehily, A.M., 2007. Milk and dairy consumption, diabetes and the metabolic syndrome: the Caerphilly prospective study. *Journal of Epidemiology & Community Health*, *61*(8), pp.695-698.

Everett, C.J. and Matheson, E.M., 2010. Biomarkers of pesticide exposure and diabetes in the 1999–2004 National Health and Nutrition Examination Survey. *Environment international*, *36*(4), pp.398-401.

Fang, L. and Li, X., 2016. Level of serum phosphorus and adult type 2 diabetes mellitus. *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical sciences*, *41*(5), pp.502-506.

Fei, G.A.O., Pan, J.M., Hou, X.H., Fang, Q.C., Lu, H.J., Tang, J.L., Gu, H.L., Pan, Z.J., Yao, Y.H., Shen, W.Z. and Jia, W.P., 2012. Liver enzymes concentrations are closely related to prediabetes: findings of the Shanghai Diabetes Study II (SHDS II). *Biomedical and Environmental Sciences*, *25*(1), pp.30-37.

FitzGerald, M.P., Litman, H.J., Link, C.L. and McKinlay, J.B., 2007. The association of nocturia with cardiac disease, diabetes, body mass index, age and diuretic use: results from the BACH survey. *The Journal of urology*, *177*(4), pp.1385-1389.

Fonville, S., Zandbergen, A.A., Vermeer, S.E., Dippel, D.W., Koudstaal, P.J. and Den Hertog, H.M., 2013. Prevalence of prediabetes and newly diagnosed diabetes in patients with a transient ischemic attack or stroke. *Cerebrovascular diseases*, *36*(4), pp.283-289.

Franks, P.W., Hanson, R.L., Knowler, W.C., Moffett, C., Enos, G., Infante, A.M., Krakoff, J. and Looker, H.C., 2007. Childhood predictors of young-onset type 2 diabetes. *Diabetes*, *56*(12), pp.2964-2972.

Fu, D.X., Cui, H.B., Guo, N.N., Su, N., Xu, J.X. and Wang, G.Y., 2016. Prediabetes and the risk of pancreatic cancer: a meta-analysis. *Int J Clin Exp Med*, *9*(10), pp.19474-19479.

Fu, X.M., Li, N., Lin, M., Zhang, W., Cheng, X.L., Liu, M.Y., Lu, Y.H. and Li, C.L., 2015. Correlation analysis of serum calcium levels and risks of diabetes mellitus in middle-aged and elderly men. *Nan fang yi ke da xue xue bao= Journal of Southern Medical University*, *35*(10), pp.1369-1373.

Funda, D.P., Kaas, A., Bock, T., Tlaskalová-Hogenová, H. and Buschard, K., 1999. Gluten-free diet prevents diabetes in NOD mice. *Diabetes/metabolism research and reviews*, *15*(5), pp.323-327.

Gabriel, S., Mihaela, I., Angela, B. and Doru, D., 2011. Prevalence of IgA antitissue transglutaminase antibodies in children with type 1 diabetes mellitus. *Journal of clinical laboratory analysis*, *25*(3), pp.156-161.

Gagnon, C., Aimé, A. and Bélanger, C., 2017. Predictors of comorbid eating disorders and diabetes in people with type 1 and type 2 diabetes. *Canadian journal of diabetes*, *41*(1), pp.52-57.

Gao, Y.X., Man, Q., Jia, S., Li, Y., Li, L. and Zhang, J., 2017. The fasting serum triglyceride levels of elderly population with different progression stages of diabetes mellitus in China. *Journal of diabetes and its complications*, *31*(12), pp.1641-1647.

Gerstein, H.C., 1994. Cow's Milk Exposure and Type I Diabetes Mellitus: A critical overview of the clinical literature. *Diabetes care*, *17*(1), pp.13-19.

Gholi, Z., Heidari-Beni, M., Feizi, A., Iraj, B. and Askari, G., 2016. The characteristics of pre-diabetic patients associated with body composition and cardiovascular disease risk factors in the Iranian population. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, *21*.

Gisi, K., Cetinkaya, A., Ozkaya, M., Kantarceken, B., Gisi, G. and Koroglu, S., 2017. Hepatitis B and C seroprevalence in patients with diabetes mellitus and its relationship with microvascular complications. *Przeglad gastroenterologiczny*, *12*(2), p.105.

Gkrania-Klotsas, E., Ye, Z., Cooper, A.J., Sharp, S.J., Luben, R., Biggs, M.L., Chen, L.K., Gokulakrishnan, K., Hanefeld, M., Ingelsson, E. and Lai, W.A., 2010. Differential white blood cell count and type 2 diabetes: systematic review and meta-analysis of cross-sectional and prospective studies. *PloS one*, *5*(10), p.e13405.

Godsland, I.F., Walton, C., Felton, C., Proudler, A., Patel, A. and Wynn, V., 1992. Insulin resistance, secretion, and metabolism in users of oral contraceptives. *The Journal of Clinical Endocrinology & Metabolism*, *74*(1), pp.64-70.

Goldberg, D.M., Martin, J.V. and Knight, A.H., 1977. Elevation of serum alkaline phosphatase activity and related enzymes in diabetes mellitus. *Clinical biochemistry*, *10*(1), pp.8-11.

Grigoriadis, E., Fam, A.G., Starok, M. and Ang, L.C., 2000. Skeletal muscle infarction in diabetes mellitus. *The Journal of rheumatology*, *27*(4), pp.1063-1068.

Gross, L.S., Li, L., Ford, E.S. and Liu, S., 2004. Increased consumption of refined carbohydrates and the epidemic of type 2 diabetes in the United States: an ecologic assessment. *The American journal of clinical nutrition*, *79*(5), pp.774-779.

Gudul, N.E., Karabag, T., Sayin, M.R., Bayraktaroglu, T. and Aydin, M., 2017. Atrial conduction times and left atrial mechanical functions and their relation with diastolic function in prediabetic patients. *The Korean journal of internal medicine*, *32*(2), p.286.

Gulcan, E., Taser, F., Toker, A., Korkmaz, U. and Alcelik, A., 2009. Increased frequency of prediabetes in patients with irritable bowel syndrome. *The American journal of the medical sciences*, *338*(2), pp.116-119.

Guo, X., Meng, G., Liu, F., Zhang, Q., Liu, L., Wu, H., Du, H., Shi, H., Xia, Y., Liu, X. and Li, C., 2016. Serum levels of immunoglobulins in an adult population and their relationship with type 2 diabetes. *Diabetes research and clinical practice*, *115*, pp.76-82.

Gupta, S. and Wang, Z., 2016. Predictors of sleep disorders among patients with type 2 diabetes mellitus. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *10*(4), pp.213-220.

Haffner, S.M., Stern, M.P., Hazuda, H.P., Mitchell, B.D. and Patterson, J.K., 1990. Cardiovascular risk factors in confirmed prediabetic individuals: does the clock for coronary heart disease start ticking before the onset of clinical diabetes?. *Jama*, *263*(21), pp.2893-2898.

Haghighatdoost, F., Amini, M., Feizi, A. and Iraj, B., 2017. Are body mass index and waist circumference significant predictors of diabetes and prediabetes risk: Results from a population based cohort study. *World journal of diabetes*, *8*(7), p.365.

Harris, M.I., 2000. Health care and health status and outcomes for patients with type 2 diabetes. *Diabetes Care*, *23*(6), pp.754-758.

Harris, S.S. and Chew, A., 2014. Predictors of Weight Loss in African Americans with Prediabetes or Early Diabetes. *Journal of the National Medical Association*, *106*, pp.8-14.

Hashimoto, Y., Hamaguchi, M., Nakanishi, N., Ohbora, A., Kojima, T. and Fukui, M., 2017. Urinary pH is a predictor of diabetes in men; a population based large scale cohort study. *Diabetes research and clinical practice*, *130*, pp.9-14.

Hasosah, M., Bokhari, A., Alsahafi, A., Sukkar, G. and Alzaben, A., 2016. A rare association of hepatitis A virus infection with type-1 diabetes. *Clinics and practice*, *6*(2).

Heianza, Y., Hara, S., Arase, Y., Saito, K., Totsuka, K., Tsuji, H., Kodama, S., Hsieh, S.D., Yamada, N., Kosaka, K. and Sone, H., 2011. Low serum potassium levels and risk of type 2

diabetes: the Toranomon Hospital Health Management Center Study 1 (TOPICS 1). *Diabetologia*, *54*(4), pp.762-766.

Hilawe, E.H., Chiang, C., Yatsuya, H., Wang, C., Ikerdeu, E., Honjo, K., Mita, T., Cui, R., Hirakawa, Y., Madraisau, S. and Ngirmang, G., 2016. Prevalence and predictors of prediabetes and diabetes among adults in Palau: population-based national STEPS survey. *Nagoya journal of medical science*, *78*(4), p.475.

Howard, A.A., Klein, R.S. and Schoenbaum, E.E., 2003. Association of hepatitis C infection and antiretroviral use with diabetes mellitus in drug users. *Clinical Infectious Diseases*, *36*(10), pp.1318-1323.

Huang, Y., Cai, X., Qiu, M., Chen, P., Tang, H., Hu, Y. and Huang, Y., 2014. Prediabetes and the risk of cancer: a meta-analysis.

Irace, C., Carallo, C., Scavelli, F., De Franceschi, M.S., Esposito, T. and Gnasso, A., 2014. Blood viscosity in subjects with normoglycemia and prediabetes. *Diabetes care*, *37*(2), pp.488-492.

Jevrić-Čaušević, A., Malenica, M. and Dujić, T., 2006. Creatine kinase activity in patients with diabetes mellitus type I and type II. *Bosnian journal of basic medical sciences*, *6*(3), pp.5-9.

Jia, Z., Zhang, X., Kang, S. and Wu, Y., 2013. Serum uric acid levels and incidence of impaired fasting glucose and type 2 diabetes mellitus: a meta-analysis of cohort studies. *Diabetes research and clinical practice*, *101*(1), pp.88-96.

Jung, M.E., Bourne, J.E., Beauchamp, M.R., Robinson, E. and Little, J.P., 2015. High-intensity interval training as an efficacious alternative to moderate-intensity continuous training for adults with prediabetes. *Journal of diabetes research*, *2015*.

Kalyani, R.R., Ji, N., Carnethon, M., Bertoni, A.G., Selvin, E., Gregg, E.W., Sims, M. and Golden, S.H., 2017. Diabetes, depressive symptoms, and functional disability in African Americans: the Jackson Heart Study. *Journal of Diabetes and its Complications*, *31*(8), pp.1259-1265.

Karim, M.N., Ahmed, K.R., Bukht, M.S., Akter, J., Chowdhury, H.A., Hossain, S., Anwar, N., Selim, S., Chowdhury, S.H., Hossain, F. and Ali, L., 2013. Pattern and predictors of dyslipidemia in patients with type 2 diabetes mellitus. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *7*(2), pp.95-100.

Khalili, M., Lombardero, M., Chung, R.T., Terrault, N.A., Ghany, M.G., Kim, W., Lau, D., Lisker-Melman, M., Sanyal, A. and Lok, A.S., 2015. Diabetes and prediabetes in patients with hepatitis B residing in North America. *Hepatology*, *62*(5), pp.1364-1374.

Khambaty, T., Callahan, C.M., Perkins, A.J. and Stewart, J.C., 2017. Depression and Anxiety Screens as Simultaneous Predictors of 10-Year Incidence of Diabetes Mellitus in Older Adults in Primary Care. *Journal of the American Geriatrics Society*, *65*(2), pp.294-300.

Klein, B.E., Klein, R. and Moss, S.E., 1998. Self-rated health and diabetes of long duration: the Wisconsin Epidemiologic Study of Diabetic Retinopathy. *Diabetes Care*, *21*(2), pp.236-240.

Ko, S.H., Baeg, M.K., Han, K.D., Ko, S.H. and Ahn, Y.B., 2015. Increased liver markers are associated with higher risk of type 2 diabetes. *World Journal of Gastroenterology: WJG*, *21*(24), p.7478.

Kohler, I.V. and Soldo, B.J., 2005. Childhood predictors of late-life diabetes: The case of Mexico. *Social biology*, *52*(3-4), pp.112-131.

Kowall, B., Lehnich, A.T., Strucksberg, K.H., Führer, D., Erbel, R., Jankovic, N., Moebus, S., Jöckel, K.H. and Stang, A., 2016. Associations among sleep disturbances, nocturnal sleep duration, daytime napping, and incident prediabetes and type 2 diabetes: the Heinz Nixdorf Recall Study. *Sleep medicine*, *21*, pp.35-41.

Kowall, B., Rathmann, W., Strassburger, K., Heier, M., Holle, R., Thorand, B., Giani, G., Peters, A. and Meisinger, C., 2010. Association of passive and active smoking with incident type 2 diabetes mellitus in the elderly population: the KORA S4/F4 cohort study. *European journal of epidemiology*, *25*(6), pp.393-402.

Kristensen, S.L., Preiss, D., Jhund, P.S., Squire, I., Cardoso, J.S., Merkely, B., Martinez, F., Starling, R.C., Desai, A.S., Lefkowitz, M.P. and Rizkala, A.R., 2016. Risk Related to Pre–Diabetes Mellitus and Diabetes Mellitus in Heart Failure With Reduced Ejection Fraction CLINICAL PERSPECTIVE: Insights From Prospective Comparison of ARNI With ACEI to Determine Impact on Global Mortality and Morbidity in Heart Failure Trial. *Circulation: Heart Failure*, *9*(1), p.e002560.

Kroenke, K. and Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals*, *32*(9), pp.509-515.

Kroenke, K., Spitzer, R.L. and Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure.[Research Support. *Non-US Gov't*.

Kubo, M., Yonemoto, K., Ninomiya, T., Iwase, M., Tanizaki, Y., Shikata, K., Iida, M. and Kiyohara, Y., 2007. Liver enzymes as a predictor for incident diabetes in a Japanese population: the Hisayama study. *Obesity*, *15*(7), pp.1841-1850.

Kurihara, O., Takano, M., Yamamoto, M., Shirakabe, A., Kimata, N., Inami, T., Kobayashi, N., Munakata, R., Murakami, D., Inami, S. and Okamatsu, K., 2013. Impact of prediabetic status on coronary atherosclerosis: a multivessel angioscopic study. *Diabetes Care*, *36*(3), pp.729-733.

Lazarov, G., Danev, S., Manolov, D. and Dobrev, S., 1990. Creatine kinase in patients with diabetes mellitus. *Vutreshni bolesti*, *29*(6), pp.77-83.

Lee, J., Keam, B., Jang, E.J., Park, M.S., Lee, J.Y., Kim, D.B., Lee, C.H., Kim, T., Oh, B., Park, H.J. and Kwack, K.B., 2011. Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data. *Osong public health and research perspectives*, *2*(2), pp.75-82.

Lee, M., Saver, J.L., Hong, K.S., Song, S., Chang, K.H. and Ovbiagele, B., 2012. Effect of pre-diabetes on future risk of stroke: meta-analysis. *Bmj*, *344*, p.e3564.

Lee, S.E., Lee, Y.B., Jun, J.E., Jin, S.M., Jee, J.H., Bae, J.C. and Kim, J.H., 2017. Increment of serum bilirubin as an independent marker predicting new-onset type 2 diabetes mellitus in a Korean population. *Nutrition, Metabolism and Cardiovascular Diseases*, *27*(3), pp.234-240.

Lee, Y.S., Cho, Y., Burgess, S., Davey Smith, G., Relton, C.L., Shin, S.Y. and Shin, M.J., 2016. Serum gamma-glutamyl transferase and risk of type 2 diabetes in the general Korean population: a Mendelian randomization study. *Human molecular genetics*, *25*(17), pp.3877-3886.

Li, B., Trakarnwijitr, I., Adams, H., Layland, J., Garlick, J. and Wilson, A., 2017. High Haemoglobin and Haematocrit Are Associated with Triglycerides, Low-density Lipoprotein and High-density Lipoprotein and Negatively Associated with Diabetes, in Patients at High Risk of Coronary Artery Disease. *Heart, Lung and Circulation*, *26*, p.S117.

Liamis, G., Liberopoulos, E., Barkas, F. and Elisaf, M., 2014. Diabetes mellitus and electrolyte disorders. *World Journal of Clinical Cases: WJCC*, *2*(10), p.488.

Liccini, A.P. and Malmstrom, T.K., 2016. Frailty and sarcopenia as predictors of adverse health outcomes in persons with diabetes mellitus. *Journal of the American Medical Directors Association*, *17*(9), pp.846-851.

Lifford, K.L., Curhan, G.C., Hu, F.B., Barbieri, R.L. and Grodstein, F., 2005. Type 2 diabetes mellitus and risk of developing urinary incontinence. Journal of the American Geriatrics Society, 53(11), pp.1851-1857.

Lin, K.D., Lee, M.Y., Feng, C.C., Chen, B.K., Yu, M.L. and Shin, S.J., 2014. Residual effect of reductions in red blood cell count and haematocrit and haemoglobin levels after 10-month withdrawal of pioglitazone in patients with Type 2 diabetes. *Diabetic Medicine*, *31*(11), pp.1341-1349.

Liu, Y., Li, J., Zhang, Z., Tang, Y., Chen, Z. and Wang, Z., 2013. Effects of exercise intervention on vascular endothelium functions of patients with impaired glucose tolerance during prediabetes mellitus. *Experimental and therapeutic medicine*, *5*(6), pp.1559-1565.

Longo-Mbenza, B., Kasiam Lasi On'kin, J.B., Nge Okwe, A., Kangola Kabangu, N. and Mbungu Fuele, S., 2010. Metabolic syndrome, aging, physical inactivity, and incidence of type 2 diabetes in general African population. *Diabetes and Vascular Disease Research*, *7*(1), pp.28-39.

Lorga, T., Aung, M.N., Naunboonruang, P., Thinuan, P., Praipaksin, N., Deesakul, T., Inwan, U., Yingtaweesak, T., Manokulanan, P., Suangkaew, S. and Payaprom, A., 2012. Predicting prediabetes in a rural community: a survey among the Karen ethnic community, Thasongyang, Thailand. *International journal of general medicine*, *5*, p.219.

Ludvigsson, J.F., Ludvigsson, J., Ekbom, A. and Montgomery, S.M., 2006. Celiac disease and risk of subsequent type 1 diabetes: a general population cohort study of children and adolescents. *Diabetes care*, *29*(11), pp.2483-2488.

Luo, J., Manson, J.E., Urrutia, R.P., Hendryx, M., LeBlanc, E.S. and Margolis, K.L., 2017. Risk of Diabetes After Hysterectomy With or Without Oophorectomy in Postmenopausal Women. *American journal of epidemiology*, *185*(9), pp.777-785.

Lyssenko, V., Almgren, P., Anevski, D., Perfekt, R., Lahti, K., Nissén, M., Isomaa, B., Forsen, B., Homström, N., Saloranta, C. and Taskinen, M.R., 2005. Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. *Diabetes*, *54*(1), pp.166-174.

Mainous, A.G., Tanner, R.J., Baker, R., Zayas, C.E. and Harle, C.A., 2014. Prevalence of prediabetes in England from 2003 to 2011: population-based, cross-sectional study. *BMJ open*, *4*(6), p.e005002.

Majidiani, H., Dalvand, S., Daryani, A., Galvan-Ramirez, M.D.L.L. and Foroutan-Rad, M., 2016. Is chronic toxoplasmosis a risk factor for diabetes mellitus? A systematic review and meta-analysis of case-control studies. *Brazilian Journal of Infectious Diseases*, *20*(6), pp.605-609.

Malin, S.K., Gerber, R., Chipkin, S.R. and Braun, B., 2012. Independent and combined effects of exercise training and metformin on insulin sensitivity in individuals with prediabetes. *Diabetes care*, *35*(1), pp.131-136.

Malin, S.K., Solomon, T.P., Blaszczak, A., Finnegan, S., Filion, J. and Kirwan, J.P., 2013. Pancreatic β-cell function increases in a linear dose-response manner following exercise training in adults with prediabetes. *American Journal of Physiology-Endocrinology and Metabolism*, *305*(10), pp.E1248-E1254.

Mamtani, M., Kulkarni, H., Wong, G., Weir, J.M., Barlow, C.K., Dyer, T.D., Almasy, L., Mahaney, M.C., Comuzzie, A.G., Glahn, D.C. and Magliano, D.J., 2016. Lipidomic risk score independently and cost-effectively predicts risk of future type 2 diabetes: results from diverse cohorts. *Lipids in health and disease*, *15*(1), p.67.

Marini, M.A., Fiorentino, T.V., Andreozzi, F., Mannino, G.C., Succurro, E., Sciacqua, A., Perticone, F. and Sesti, G., 2017. Hemorheological alterations in adults with prediabetes identified by hemoglobin A1c levels. *Nutrition, Metabolism and Cardiovascular Diseases*, *27*(7), pp.601-608.

Mathenge, N., Fan, W., Wong, N.D., Hirsch, C., Delaney, J.A., Amsterdam, E.A., Koch, B., Calara, R. and Gardin, J.M., 2017. Pre-Diabetes, Diabetes and Predictors of Incident Angina During Long-term Evaluation in Older Women and Men in the Cardiovascular Health Study.

Maxwell, D.B., Fisher, E.A., Ross-Clunis 3rd, H.A. and Estep, H.L., 1986. Serum alkaline phosphatase in diabetes mellitus. *Journal of the American College of Nutrition*, *5*(1), pp.55-59.

Mayans, L., 2015. Metabolic syndrome: insulin resistance and prediabetes. *FP essentials*, *435*, pp.11-16.

McFarland, K.F., Catalano, E.W., Day, J.F., Thorpe, S.R. and Baynes, J.W., 1979. Nonenzymatic glucosylation of serum proteins in diabetes mellitus. *Diabetes*, *28*(11), pp.1011-1014.

McMillan, D.E., 1989. Increased levels of acute-phase serum proteins in diabetes. *Metabolism-Clinical and Experimental*, *38*(11), pp.1042-1046.

Meisinger, C., Rückert, I.M., Stöckl, D., Thorand, B., Peters, A., Kowall, B. and Rathmann, W., 2014. Hematological parameters and prediabetes and diabetes in adults from the general population: a cross-sectional study. *J Diabetes Metab*, *5*(335), p.2.

Meisinger, C., Stöckl, D., Rückert, I.M., Döring, A., Thorand, B., Heier, M., Huth, C., Belcredi, P., Kowall, B. and Rathmann, W., 2013. Serum potassium is associated with prediabetes and newly diagnosed diabetes in hypertensive adults from the general population: the KORA F4-study. *Diabetologia*, *56*(3), pp.484-491.

Melsom, T., Schei, J., Stefansson, V.T.N., Solbu, M.D., Jenssen, T.G., Mathisen, U.D., Wilsgaard, T. and Eriksen, B.O., 2016. Prediabetes and risk of glomerular hyperfiltration and albuminuria in the general nondiabetic population: a prospective cohort study. *American Journal of Kidney Diseases*, *67*(6), pp.841-850.

Micha, R., Wallace, S.K. and Mozaffarian, D., 2010. Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. *Circulation*, *121*(21), pp.2271-2283.

Montgomery, M.P., Kamel, F., Saldana, T.M., Alavanja, M.C.R. and Sandler, D.P., 2008. Incident diabetes and pesticide exposure among licensed pesticide applicators: Agricultural Health Study, 1993–2003. *American journal of epidemiology*, *167*(10), pp.1235-1246.

Mukhtar, N.A., Ayala, C., Maher, J.J. and Khalili, M., 2012. Assessment of factors associated with pre-diabetes in HCV infection including direct and dynamic measurements of insulin action. *Journal of viral hepatitis*, *19*(7), pp.480-487.

Musselman, D.L., Betan, E., Larsen, H. and Phillips, L.S., 2003. Relationship of depression to diabetes types 1 and 2: epidemiology, biology, and treatment. *Biological psychiatry*, *54*(3), pp.317-329.

Musumeci, V., Cherubim, P., Zuppi, C., Zappacosta, B., Ghirlanda, G. and Salvo, S., 1993. Aminotransferases and lactate dehydrogenase in saliva of diabetic patients. *Journal of oral pathology & medicine*, *22*(2), pp.73-76.

Naas, A.A., Davidson, N.C., Thompson, C., Cummings, F., Ogston, S.A., Jung, R.T., Newton, R.W. and Struthers, A.D., 1998. QT and QTc dispersion are accurate predictors of cardiac death in newly diagnosed non-insulin dependent diabetes: cohort study. *Bmj*, *316*(7133), pp.745-746.

Nakanishi, N., Suzuki, K. and Tatara, K., 2004. Haematocrit and risk of development of Type 2 diabetes mellitus in middle-aged Japanese men. *Diabetic medicine*, *21*(5), pp.476-482.

Nelson, K.M., Chapko, M.K., Reiber, G. and Boyko, E.J., 2005. The association between health insurance coverage and diabetes care; data from the 2000 Behavioral Risk Factor Surveillance System. *Health services research*, *40*(2), pp.361-372.

Nguyen, Q.M., Srinivasan, S.R., Xu, J.H., Chen, W., Hassig, S., Rice, J. and Berenson, G.S., 2011. Elevated liver function enzymes are related to the development of prediabetes and type 2 diabetes in younger adults: the Bogalusa Heart Study. *Diabetes care*, *34*(12), pp.2603-2607.

Nilsson, P.M., Theobald, H., Journath, G. and Fritz, T., 2004. Gender differences in risk factor control and treatment profile in diabetes: a study in 229 Swedish primary health care centres. *Scandinavian journal of primary health care*, *22*(1), pp.27-31.

Nishi, T., Babazono, A., Maeda, T., Imatoh, T. and Une, H., 2015. Evaluation of the fatty liver index as a predictor for the development of diabetes among insurance beneficiaries with prediabetes. *Journal of diabetes investigation*, *6*(3), pp.309-316.

Nouwen, A., Ford, T., Balan, A.T., Twisk, J., Ruggiero, L. and White, D., 2011. Longitudinal motivational predictors of dietary self-care and diabetes control in adults with newly diagnosed type 2 diabetes mellitus. *Health Psychology*, *30*(6), p.771.

Okada, R., Yasuda, Y., Tsushita, K., Wakai, K., Hamajima, N. and Matsuo, S., 2011. Glomerular hyperfiltration in prediabetes and prehypertension. *Nephrology Dialysis Transplantation*, *27*(5), pp.1821-1825.

Okwechime, I.O. and Roberson, S., 2015. Prevalence and predictors of pre-diabetes and diabetes among adults 18 years or older in Florida: A multinomial logistic modeling approach. *PloS one*, *10*(12), p.e0145781.

Origuchi, T., Yamaguchi, S., Inoue, A., Kazaura, Y., Matsuo, N., Abiru, N., Kawakami, A. and Eguchi, K., 2011. Increased incidence of pre-diabetes mellitus at a department of rheumatology: a retrospective study. *Modern rheumatology*, *21*(5), pp.495-499.

Östenson, C.G., Hilding, A., Grill, V. and Efendic, S., 2012. High consumption of smokeless tobacco ("snus") predicts increased risk of type 2 diabetes in a 10-year prospective study of middle-aged Swedish men. *Scandinavian journal of public health*, *40*(8), pp.730-737.

Palmer, B.F. and Clegg, D.J., 2015. Electrolyte and acid–base disturbances in patients with diabetes mellitus. *New England Journal of Medicine*, *373*(6), pp.548-559.

Pan, A., Wang, Y., Talaei, M., Hu, F.B. and Wu, T., 2015. Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. *The lancet Diabetes & endocrinology*, *3*(12), pp.958-967.

Parildar, H., Gulmez, O., Cigerli, O., Unal, A.D., Erdal, R. and Demirag, N.G., 2013. Carotid Artery Intima Media Thickness and HsCRP; Predictors for Atherosclerosis in Prediabetic Patients?. *Pakistan journal of medical sciences*, *29*(2), p.495.

Pereira, R.R., Amladi, S.T. and Varthakavi, P.K., 2011. A study of the prevalence of diabetes, insulin resistance, lipid abnormalities, and cardiovascular risk factors in patients with chronic plaque psoriasis. *Indian journal of dermatology*, *56*(5), p.520.

Perreault, L., Ma, Y., Dagogo-Jack, S., Horton, E., Marrero, D., Crandall, J. and Barrett-Connor, E., 2008. Sex differences in diabetes risk and the effect of intensive lifestyle modification in the Diabetes Prevention Program. *Diabetes care*, *31*(7), pp.1416-1421.

Perreault, L., Pan, Q., Mather, K.J., Watson, K.E., Hamman, R.F., Kahn, S.E. and Diabetes Prevention Program Research Group, 2012. Effect of regression from prediabetes to normal glucose regulation on long-term reduction in diabetes risk: results from the Diabetes Prevention Program Outcomes Study. *The Lancet*, *379*(9833), pp.2243-2251.

Perros, P., McCrimmon, R.J., Shaw, G. and Frier, B.M., 1995. Frequency of thyroid dysfunction in diabetic patients: value of annual screening. *Diabetic medicine*, *12*(7), pp.622-627.

Plantinga, L.C., Crews, D.C., Coresh, J., Miller, E.R., Saran, R., Yee, J., Hedgeman, E., Pavkov, M., Eberhardt, M.S., Williams, D.E. and Powe, N.R., 2010. Prevalence of chronic kidney disease in US adults with undiagnosed diabetes or prediabetes. *Clinical Journal of the American Society of Nephrology*, pp.CJN-07891109.

Portnoy, D.B., Kaufman, A.R., Klein, W.M., Doyle, T.A. and De Groot, M., 2014. Cognitive and affective perceptions of vulnerability as predictors of exercise intentions among people with type 2 diabetes. *Journal of risk research*, *17*(2), pp.177-193.

Qureshi, A.A., Choi, H.K., Setty, A.R. and Curhan, G.C., 2009. Psoriasis and the risk of diabetes and hypertension: a prospective study of US female nurses. *Archives of dermatology*, *145*(4), pp.379-382.

Radaideh, A.R.M., Nusier, M.K., Amari, F.L., Bateiha, A.E., El-Khateeb, M.S., Naser, A.S. and Ajlouni, K.M., 2004. Thyroid dysfunction in patients with type 2 diabetes mellitus in Jordan. *Saudi medical journal*, *25*(8), pp.1046-1050.

Rapoport, M.J., Bistritzer, T., Vardi, O., Broide, E., Azizi, A. and Vardi, P., 1996. Increased prevalence of diabetes-related autoantibodies in celiac disease. *Journal of pediatric gastroenterology and nutrition*, *23*(5), pp.524-527.

Risérus, U., Ärnlöv, J., Brismar, K., Zethelius, B., Berglund, L. and Vessby, B., 2004. Sagittal abdominal diameter is a strong anthropometric marker of insulin resistance and hyperproinsulinemia in obese men. *Diabetes care*, *27*(8), pp.2041-2046.

Rooney, M.R., Pankow, J.S., Sibley, S.D., Selvin, E., Reis, J.P., Michos, E.D. and Lutsey, P.L., 2016. Serum calcium and incident type 2 diabetes: the Atherosclerosis Risk in Communities (ARIC) study, 2. *The American journal of clinical nutrition*, *104*(4), pp.1023-1029.

Sacerdote, C., Ricceri, F., Rolandsson, O., Baldi, I., Chirlaque, M.D., Feskens, E., Bendinelli, B., Ardanaz, E., Arriola, L., Balkau, B. and Bergmann, M., 2012. Lower educational level is a predictor of incident type 2 diabetes in European countries: the EPIC-InterAct study. *International journal of epidemiology*, *41*(4), pp.1162-1173.

Saito, T., Ishikawa, S.E., Higashiyama, M., Nakamura, T., Rokkaku, K., Hayashi, H., Kusaka, I., Nagasaka, S. and Saito, T., 1999. Inverse distribution of serum sodium and potassium in uncontrolled inpatients with diabetes mellitus. *Endocrine journal*, *46*(1), pp.75-80.

Salazar Vázquez, B.Y., Salazar Vázquez, M.A., Jáquez, M.G., Bracho Huemoeller, A.H., Intaglietta, M. and Cabrales, P., 2010. Blood pressure directly correlates with blood viscosity in diabetes type 1 children but not in normals. *Clinical hemorheology and microcirculation*, *44*(1), pp.55-61.

Salomaa, V., Havulinna, A., Saarela, O., Zeller, T., Jousilahti, P., Jula, A., Muenzel, T., Aromaa, A., Evans, A., Kuulasmaa, K. and Blankenberg, S., 2010. Thirty-one novel biomarkers as predictors for clinically incident diabetes. *PloS one*, *5*(4), p.e10100.

Salomaa, V., Riley, W., Kark, J.D., Nardo, C. and Folsom, A.R., 1995. Non–insulin-dependent diabetes mellitus and fasting glucose and insulin concentrations are associated with arterial stiffness indexes: the ARIC Study. *Circulation*, *91*(5), pp.1432-1443.

Satman, I., Omer, B., Tutuncu, Y., Kalaca, S., Gedik, S., Dinccag, N., Karsidag, K., Genc, S., Telci, A., Canbaz, B. and Turker, F., 2013. Twelve-year trends in the prevalence and risk factors of diabetes and prediabetes in Turkish adults. *European journal of epidemiology*, *28*(2), pp.169-180.

Schootman, M., Andresen, E.M., Wolinsky, F.D., Malmstrom, T.K., Miller, J.P., Yan, Y. and Miller, D.K., 2007. The effect of adverse housing and neighborhood conditions on the

development of diabetes mellitus among middle-aged African Americans. *American journal of epidemiology*, *166*(4), pp.379-387.

Seligman, H.K., Bindman, A.B., Vittinghoff, E., Kanaya, A.M. and Kushel, M.B., 2007. Food insecurity is associated with diabetes mellitus: results from the National Health Examination and Nutrition Examination Survey (NHANES) 1999–2002. *Journal of general internal medicine*, *22*(7), pp.1018-1023.

Shamshirgaran, S.M., Ataei, J., Alamdari, M.I., Safaeian, A. and Aminisani, N., 2016. Predictors of health-related quality of life among people with type II diabetes Mellitus in Ardabil, Northwest of Iran, 2014. *Primary care diabetes*, *10*(4), pp.244-250.

Shapiro, J., Cohen, A.D., David, M., Hodak, E., Chodik, G., Viner, A., Kremer, E. and Heymann, A., 2007. The association between psoriasis, diabetes mellitus, and atherosclerosis in Israel: a case-control study. *Journal of the American Academy of Dermatology*, *56*(4), pp.629-634.

Shearer, D.M., Thomson, W.M., Broadbent, J.M., McLean, R., Poulton, R. and Mann, J., 2016. High-risk glycated hemoglobin trajectories established by mid-20s: findings from a birth cohort study. *BMJ Open Diabetes Research and Care*, *4*(1), p.e000243.

Simmons, D., 2010. Increased red cell count in diabetes and pre-diabetes. *Diabetes research and clinical practice*, *90*(3), pp.e50-e53.

Sing, C.W., Cheng, V.K., Ho, D.K., Kung, A.W., Cheung, B.M.Y., Wong, I.C.K., Tan, K.C.B., Salas-Salvadó, J., Becerra-Tomas, N. and Cheung, C.L., 2016. Serum calcium and incident diabetes: an observational study and meta-analysis. *Osteoporosis International*, *27*(5), pp.1747-1754.

Singh, K.K., Panda, S.K. and Acharya, S.K., 2013. Patients with diabetes mellitus are prone to develop severe hepatitis and liver failure due to hepatitis virus infection. *Journal of clinical and experimental hepatology*, *3*(4), pp.275-280.

Soewondo, P. and Pramono, L.A., 2011. Prevalence, characteristics, and predictors of pre-diabetes in Indonesia. *Medical Journal of Indonesia*, *20*(4), p.283.

Spitzer, R.L., Kroenke, K., Williams, J.B. and Patient Health Questionnaire Primary Care Study Group, 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama*, *282*(18), pp.1737-1744.

Stene, L.C. and Nafstad, P., 2001. Relation between occurrence of type 1 diabetes and asthma. *The Lancet*, *357*(9256), pp.607-608.

Stojkovikj, J., Zafirova-Ivanovska, B., Kaeva, B., Anastasova, S., Angelovska, I., Jovanovski, S. and Stojkovikj, D., 2016. The Prevalence of diabetes mellitus in COPD patients with severe and

very severe stage of the disease. *Open access Macedonian journal of medical sciences*, *4*(2), p.253.

Suh, S., Bae, J.C., Jin, S.M., Jee, J.H., Park, M.K., Kim, D.K. and Kim, J.H., 2017. Serum calcium changes and risk of type 2 diabetes mellitus in Asian population. *diabetes research and clinical practice*, *133*, pp.109-114.

Tabá, G., Herder, C., Rathmann, W., Brunner, E. and Kivimäk, M., 2012. Prediabetes: A high-risk state for developing diabetes. *Lancet*, *379*(9833), pp.2279-2290.

Tamariz, L.J., Young, J.H., Pankow, J.S., Yeh, H.C., Schmidt, M.I., Astor, B. and Brancati, F.L., 2008. Blood viscosity and hematocrit as risk factors for type 2 diabetes mellitus: the atherosclerosis risk in communities (ARIC) study. *American journal of epidemiology*, *168*(10), pp.1153-1160.

Taniguchi, A., Fukushima, M., Seino, Y., Sakai, M., Yoshii, S., Nagasaka, S., Yamauchi, I., Okumura, T., Nin, K., Tokuyama, K. and Yamadori, N., 2003. Platelet count is independently associated with insulin resistance in non-obese Japanese type 2 diabetic patients. *Metabolism-Clinical and Experimental*, *52*(10), pp.1246-1249.

Tawfik, S.H., Mahmoud, B.F., Saad, M.I., Shehata, M., Kamel, M.A. and Helmy, M.H., 2015. Similar and additive effects of ovariectomy and diabetes on insulin resistance and lipid metabolism. *Biochemistry research international*, *2015*.

Tester, J., Sharma, S., Jasik, C.B., Mietus-Snyder, M. and Tinajero-Deck, L., 2013. Gender differences in prediabetes and insulin resistance among 1356 obese children in Northern California. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *7*(3), pp.161-165.

Thomas, M.C., MacIsaac, R.J., Tsalamandris, C., Power, D. and Jerums, G., 2003. Unrecognized anemia in patients with diabetes: a cross-sectional survey. *Diabetes care*, *26*(4), pp.1164-1169.

Toker, S., Shirom, A., Melamed, S. and Armon, G., 2012. Work characteristics as predictors of diabetes incidence among apparently healthy employees. *Journal of occupational health psychology*, *17*(3), p.259.

Tsenkova, V., Pudrovska, T. and Karlamangla, A., 2014. Childhood Socioeconomic Disadvantage and Pre-diabetes and Diabetes in Later Life: A Study of Biopsychosocial Pathways. *Psychosomatic medicine*, *76*(8), p.622.

Tulloch-Reid, M.K., Hanson, R.L., Saremi, A., Looker, H.C., Williams, D.E., Krakoff, J. and Knowler, W.C., 2004. Hematocrit and the incidence of type 2 diabetes in the Pima Indians. *Diabetes Care*, *27*(9), pp.2245-2246.

Tuomainen, T.P., Nyyssönen, K., Salonen, R., Tervahauta, A., Korpela, H., Lakka, T., Kaplan, G.A. and Salonen, J.T., 1997. Body iron stores are associated with serum insulin and blood

glucose concentrations: population study in 1,013 eastern Finnish men. *Diabetes care*, *20*(3), pp.426-428.

Twig, G., Afek, A., Shamiss, A., Derazne, E., Tzur, D., Gordon, B. and Tirosh, A., 2013. White blood cells count and incidence of type 2 diabetes in young men. *Diabetes care*, *36*(2), pp.276-282.

Valdés-Ramos, R., Ana Laura, G.L., Beatriz Elina, M.C. and Alejandra Donaji, B.A., 2015. Vitamins and type 2 diabetes mellitus. *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, *15*(1), pp.54-63.

Van Der Schaft, N., Brahimaj, A., Wen, K.X., Franco, O.H. and Dehghan, A., 2017. The association between serum uric acid and the incidence of prediabetes and type 2 diabetes mellitus: The Rotterdam Study. *PloS one*, *12*(6), p.e0179482.

Vedovato, M., Lepore, G., Coracina, A., Dodesini, A.R., Jori, E., Tiengo, A., Del Prato, S. and Trevisan, R., 2004. Effect of sodium intake on blood pressure and albuminuria in Type 2 diabetic patients: the role of insulin resistance. *Diabetologia*, *47*(2), pp.300-303.

Venkataramana, G., Indira, P. and Rao, D.V.M., 2013. Changes of plasma total proteins, Albumin and Fibrinogen in Type 2 Diabetes mellitus-A Pilot study. *Original article*, *7*(2), pp.679-685.

Veronese, N., Stubbs, B., Fontana, L., Trevisan, C., Bolzetta, F., De Rui, M., Sartori, L., Musacchio, E., Zambon, S., Maggi, S. and Perissinotto, E., 2016. Frailty is associated with an increased risk of incident type 2 diabetes in the elderly. *Journal of the American Medical Directors Association*, *17*(10), pp.902-907.

Vesely, D.L., Dilley, R.W., Duckworth, W.C. and Paustian, F.F., 1999. Hepatitis A-induced diabetes mellitus, acute renal failure, and liver failure. *The American journal of the medical sciences*, *317*(6), pp.419-424.

Vozarova, B., Weyer, C., Lindsay, R.S., Pratley, R.E., Bogardus, C. and Tataranni, P.A., 2002. High white blood cell count is associated with a worsening of insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes*, *51*(2), pp.455-461.

Vučak, J., Katić, M., Bielen, I., Vrdoljak, D., Lalić, D.I., Kranjčević, K. and Marković, B.B., 2012. Association between hyperuricemia, prediabetes, and prehypertension in the Croatian adult population-a cross-sectional study. *BMC cardiovascular disorders*, *12*(1), p.117.

Walter, R.E., Beiser, A., Givelber, R.J., O'connor, G.T. and Gottlieb, D.J., 2003. Association between glycemic state and lung function: the Framingham Heart Study. *American journal of respiratory and critical care medicine*, *167*(6), pp.911-916.

Wang, C.S., Wang, S.T., Yao, W.J., Chang, T.T. and Chou, P., 2003. Community-based study of hepatitis C virus infection and type 2 diabetes: an association affected by age and hepatitis severity status. *American journal of epidemiology*, *158*(12), pp.1154-1160.

Wang, H., Shara, N.M., Calhoun, D., Umans, J.G., Lee, E.T. and Howard, B.V., 2010. Incidence rates and predictors of diabetes in those with prediabetes: the Strong Heart Study. *Diabetes/metabolism research and reviews*, *26*(5), pp.378-385.

Wang, H., Song, Y., Sun, S., Gao, L., Liu, L., Meng, G., Wu, H., Xia, Y., Bao, X., Gu, Y. and Shi, H., 2017. The association between immunoglobulin concentrations and prediabetes prevalence in a large Chinese cohort. *Metabolism-Clinical and Experimental*, *73*, pp.77-84.

Wang, J., Li, Y., Han, X., Hu, H., Wang, F., Li, X., Yang, K., Yuan, J., Yao, P., Miao, X. and Wei, S., 2017. Serum bilirubin levels and risk of type 2 diabetes: results from two independent cohorts in middle-aged and elderly Chinese. *Scientific reports*, *7*, p.41338.

Wang, Z.S., Song, Z.C., Bai, J.H., Li, F., Wu, T., Qi, J. and Hu, J., 2013. Red blood cell count as an indicator of microvascular complications in Chinese patients with type 2 diabetes mellitus. *Vascular health and risk management*, *9*, p.237.

West, B., Luke, A., Durazo-Arvizu, R.A., Cao, G., Shoham, D. and Kramer, H., 2008. Metabolic syndrome and self-reported history of kidney stones: the National Health and Nutrition Examination Survey (NHANES III) 1988-1994. *American Journal of Kidney Diseases*, *51*(5), pp.741-747.

Wilson, L. and Mishra, G., 2017. Quantifying the mediating effect of body mass index on the association between hysterectomy status and incident diabetes in a mid-aged cohort of Australian women. *Maturitas*, *100*, pp.136-137.

Winnard, D., Wright, C., Jackson, G., Gow, P., Kerr, A., McLachlan, A., Orr-Walker, B. and Dalbeth, N., 2013. Gout, diabetes and cardiovascular disease in the Aotearoa New Zealand adult population: co-prevalence and implications for clinical practice. *The New Zealand Medical Journal (Online)*, *126*(1368).

World Health Organization, 2011. Waist circumference and waist-hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008.

Yamane, T., Yokoyama, A., Kitahara, Y., Miyamoto, S., Haruta, Y., Hattori, N., Yamane, K., Hara, H. and Kohno, N., 2013. Cross-sectional and prospective study of the association between lung function and prediabetes. *BMJ open*, *3*(2), p.e002179.

Yokota, N., Miyakoshi, T., Sato, Y., Nakasone, Y., Yamashita, K., Imai, T., Hirabayashi, K., Koike, H., Yamauchi, K. and Aizawa, T., 2017. Predictive models for conversion of prediabetes to diabetes. *Journal of diabetes and its complications*, *31*(8), pp.1266-1271.

Yong, M., Parkinson, K., Goenka, N. and O'Mahony, C., 2010. Diabetes and genital warts: an unhappy coalition. *International journal of STD & AIDS*, *21*(7), pp.457-459.

Yoshimura, K., Terada, N., Matsui, Y., Terai, A., Kinukawa, N. and Arai, Y., 2004. Prevalence of and risk factors for nocturia: analysis of a health screening program. *International journal of urology*, *11*(5), pp.282-287.

Zaccardi, F., Webb, D.R., Carter, P., Pitocco, D., Khunti, K., Davies, M.J., Kurl, S. and Laukkanen, J.A., 2015. Association between direct measurement of active serum calcium and risk of type 2 diabetes mellitus: a prospective study. *Nutrition, Metabolism and Cardiovascular Diseases*, *25*(6), pp.562-568.

Zelber-Sagi, S., Lotan, R., Shibolet, O., Webb, M., Buch, A., Nitzan-Kaluski, D., Halpern, Z., Santo, E. and Oren, R., 2013. Non-alcoholic fatty liver disease independently predicts prediabetes during a 7-year prospective follow-up. *Liver International*, *33*(9), pp.1406-1412.

Zhang, L., Zhang, Z., Zhang, Y., Hu, G. and Chen, L., 2014. Evaluation of Finnish Diabetes Risk Score in screening undiagnosed diabetes and prediabetes among US adults by gender and race: NHANES 1999-2010. *PloS one*, *9*(5), p.e97865.

Zhang, Q., Bao, X., Meng, G., Liu, L., Wu, H., Du, H., Shi, H., Xia, Y., Guo, X., Liu, X. and Li, C., 2016. The predictive value of mean serum uric acid levels for developing prediabetes. *Diabetes research and clinical practice*, *118*, pp.79-89.

Zhang, S., Liu, Y., Wang, G., Xiao, X., Gang, X., Li, F., Sun, C., Gao, Y. and Wang, G., 2016. The relationship between alcohol consumption and incidence of glycometabolic abnormality in middle-aged and elderly Chinese men. *International journal of endocrinology*, *2016*.

Zhu, H., Wang, N., Han, B., Li, Q., Chen, Y., Zhu, C., Chen, Y., Xia, F., Cang, Z., Lu, M. and Chen, C., 2016. Low Sex Hormone-Binding Globulin Levels Associate with Prediabetes in Chinese Men Independent of Total Testosterone. *PloS one*, *11*(9), p.e0162004.

Ziaee, A., Ghorbani, A., Kalbasi, S., Hejrati, A. and Moradi, S., 2017. Association of hematological indices with prediabetes: A cross-sectional study. *Electronic physician*, *9*(9), p.5206.

Zyriax, B.C., Salazar, R., Hoeppner, W., Vettorazzi, E., Herder, C. and Windler, E., 2013. The association of genetic markers for type 2 diabetes with prediabetic status-cross-sectional data of a diabetes prevention trial. *PloS one*, *8*(9), p.e75807.

Table 4: Distribution of general (socio-economic, clinical, physiological and biochemical) variables in original and multiply-imputed datasets prepared for the objective 1 and post-hoc accuracy analysis of the multiple imputation algorithm. Distributions of numeric and categorical variables between original and imputed samples were compared using two-samples t-test and chi-squared test for two proportions, respectively.

Numeric variables			
variable	**Original sample mean; SD (N)**	**Imputed sample mean; SD (N)**	**p-value**
Household income (annual)	8.35; 4.06 (6058)	8.32; 4.05 (6346)	0.6804
Family income (annual)	8.02; 4.13 (6069)	8.01; 4.12 (6346)	0.8926
Income poverty ratio	2.39; 1.65 (5854)	2.44; 1.65 (6346)	0.0945
Healthcare seeking frequency	2.34; 2.02 (6339)	2.34; 2.02 (6346)	1.0000
Number of rooms	5.99; 2.25 (6266)	5.99; 2.26 (6346)	1.0000
Monthly family income	6.94; 3.33 (5655)	6.96; 3.31 (6346)	0.7418
Poverty index	2.23; 1.59 (5655)	2.24; 1.58 (6346)	0.7300
Urinary leakage frequency	1.66; 1.17 (4513)	1.62; 1.12(6346)	0.0718
Nightly urinating frequency	1.15; 1.13 (4515)	1.15; 1.14 (6346)	1.0000
Duration of sedentary activity	439.98; 193.84 (6284)	440.08; 193.88(6346)	0.9769
Duration of watching TV	2.34; 1.63 (6309)	2.34; 1.63 (6346)	1.0000
Duration of using computer	1.17; 1.54 (6312)	1.17; 1.54 (6346)	1.0000
Duration of sleep	6.95; 1.41 (5631)	6.96; 1.41 (6346)	0.6985
Number of smokers in the household	0.39; 0.74 (6278)	0.39; 0.74 (6346)	1.0000
Pulse rate	73.30; 11.95 (6131)	73.33; 11.97 (6346)	0.8886
BMI	27.73; 7.11 (6275)	27.73; 7.09(6346)	1.0000
Arm circumference	32.00; 5.51 (6071)	32.04; 5.53 (6346)	0.6865
Waist circumference	94.43; 17.31 (6014)	94.59; 17.40 (6346)	0.6085
Sagittal abdominal diameter	21.43; 4.51 (5857)	21.53; 4.58 (6346)	0.2248
Toxoplasmosis IgG	13.18; 35.74 (5228)	13.19; 35.89 (6346)	0.9881
Toxoplasmosis IgM	0.23; 0.24 (5055)	0.23; 0.26 (6346)	1.0000
Urinary albumin	29.73; 142.45 (6198)	29.80; 141.36 (6346)	0.9780
Urinary creatinine	126.30; 81.30 (6198)	126.78; 81.44 (6346)	0.7412
Urinary albumin: creatinine ratio	28.11; 167.93 (6198)	28.16; 166.67 (6346)	0.9866
HDL	53.35; 15.59 (5949)	53.30; 15.56 (6346)	0.8588
Cholesterol	182.82; 40.51 (5949)	182.86; 40.71 (6346)	0.9565
WBC count	7.23; 2.29 (6014)	7.22; 2.27 (6346)	0.8074
Lymphocyte count	2.19; 0.94 (5996)	2.19; 0.92 (6346)	1.0000
Monocyte count	0.58; 0.19 (5996)	0.58; 0.20 (6346)	1.0000
Neutrophil count	4.21; 1.78 (5996)	4.21; 1.76 (6346)	1.0000
Eosinophil count	0.20; 0.18 (5996)	0.20; 0.18 (6346)	1.0000
Basophil count	0.04; 0.05 (5996)	0.04; 0.05 (6346)	1.0000
RBC count	4.68; 0.49 (6014)	4.68; 0.49 (6346)	1.0000
Hemoglobin	13.97; 1.49 (6014)	13.97; 1.49 (6346)	1.0000
Platelet count	238.87; 58.68 (6014)	238.95; 58.46 (6346)	0.9395
Alkaline phosphatase	79.00; 53.48 (5928)	79.41; 54.83 (6346)	0.6753
AMT	24.90; 17.85 (5927)	24.87; 17.54 (6346)	0.9252
ALT	23.68; 18.35 (5927)	23.63; 18.19 (6346)	0.8796
Serum bicarbonate	25.17; 2.24 (5929)	25.17; 2.24 (6346)	1.0000
Serum calcium	9.49; 0.36 (5894)	9.49; 0.36 (6346)	1.0000
Creatine phosphokinase	155.54; 190.26 (5918)	154.21; 186.34(6346)	0.6958
Serum chloride	104.47; 2.62 (5929)	104.48; 2.61 (6346)	0.8323
Serum creatinine	0.86; 0.40 (5929)	0.86; 0.39 (6346)	1.0000
Serum globulin	2.81; 0.43 (5924)	2.81; 0.43 (6346)	1.0000
GGT	24.30; 32.95 (5928)	24.34; 33.51 (6346)	0.9469

Serum iron	84.15; 36.81 (5905)	84.16; 36.98 (6346)	0.9880
Serum potassium	4.02; 0.35 (5928)	4.02; 0.35 (6346)	1.0000
Lactate dehydrogenase	126.38; 28.87 (5927)	126.22; 29.06 (6346)	0.7598
Sodium	139.87; 2.14 (5929)	139.86; 2.13 (6346)	0.7954
Osmolality	278.86; 4.72 (5929)	278.84; 4.71 (6346)	0.8143
Phosphorus	3.95; 0.65 (5929)	3.95; 0.65 (6346)	1.0000
Bilirubin	0.64; 0.30 (5925)	0.64; 0.30 (6346)	1.0000
Total protein	7.11; 0.46 (5924)	7.11; 0.46 (6346)	1.0000
Triglycerides	134.48; 98.07 (5926)	134.27; 97.56 (6346)	0.9054
Serum uric acid	5.31; 1.38 (5928)	5.31; 1.38 (6346)	1.0000
Vitamin B12	617.07; 510.70 (4832)	618.91; 523.53(6346)	0.8524
Processed food expenditure	0.23; 0.20 (6244)	0.24; 0.21 (6346)	**0.0062**
Number of processed food meals	9.51; 14.18 (6325)	9.52; 14.17 (6346)	0.9683
Hematocrit	41.30; 4.04 (6014)	41.34; 4.04 (6346)	0.5822
Mean SBP	119.32; 17.20 (6118)	119.32; 17.28 (6346)	1.0000
Mean DBP	66.92; 12.97 (6118)	66.85; 13.09 (6346)	0.7643
Categorical variables			
Variable	**Original sample proportion**	**Imputed sample proportion**	**p-value**
Diagnosed DM risk			
No	5254/6061	5514/6346	0.7373
Yes	807/6061	832/6346	
Self-perceived DM risk			
No	4442/6022	4678/6346	0.9525
Yes	1580/6022	1668/6346	
Blood test for DM			
No	3294/5930	3481/6346	0.4393
Yes	2636/5930	2865/6346	
Served in armed forces			
No	4995/5435	5899/6346	**0.0310**
Yes	440/5435	447/6346	
Born in the US			
USA	4793/6342	4795/6346	0.9831
Other	1549/6342	1551/6346	
Citizenship			
Yes	5573/6337	5578/6346	0.9367
No	764/6337	768/6346	
Marital status			
Married/ Living with partner	3522/4958	4627/6346	**0.0274**
Widowed/ Divorced/ Separated/ Never married	1436/4958	1719/6346	
Food security			
Full	4276/6278	4319/6346	0.9498
Marginal	708/6278	720/6346	0.9037
Low	829/6278	840/6346	0.9579
very low	465/6278	467/6346	0.9181
Health insurance			
No	1268/6338	1271/6346	0.9752
Yes	5070/6338	5075/6346	0.9752
Hepatitis B			
No	6267/6324	6289/6346	0.9851
Yes	57/6324	57/6346	0.9851
Hepatitis C			
No	6258/6324	6280/6346	0.9840
Yes	66/6324	66/6346	
Self-rated general health			
Excellent	1087/6343	1089/6346	0.9721

Very good	1771/6343	1771/6346	0.9868
Good	2317/6343	2318/6346	0.9986
Fair	975/6343	975/6346	0.9909
Poor	193/6343	193/6346	0.9962
Self-rated health trend			
Worse	615/6345	615/6346	0.9977
About the same	4454/6345	4454/6346	0.9891
Better	1276/6345	1277/6346	0.9859
Mental healthcare use			
No	5805/6342	5809/6346	0.9914
Yes	537/6342	537/6346	0.9914
Type of housing			
Owned/being bought	3685/6274	3725/6346	0.9672
Rented	2429/6274	2458/6346	0.9837
Other	160/6274	163/6346	0.9480
Hepatitis A vaccination			
Yes, at least 2 doses	3088/5538	3518/6346	0.7231
Less than 2 doses	129/5538	154/6346	0.7284
No doses	2321/5538	2674/6346	0.8031
Hepatitis B vaccination			
Yes, at least 3 doses	2786/5615	3144/6346	0.9356
Less than 3 doses	126/5615	145/6346	0.8807
No doses	2703/5615	3057/6346	0.9711
Poverty category			
Monthly poverty level index <= 1.30	2310/5991	2436/6346	0.8449
1.30 < Monthly poverty level index <= 1.85	835/5991	883/6346	0.9702
Monthly poverty level index > 1.85	2846/5991	3027/6346	0.8286
Diagnosed kidney disease			
No	4824/4955	6184/6346	0.7626
Yes	131/4955	162/6346	
Kidney stones			
No	4824/4955	5813/6346	**<0.0001**
Yes	131/4955	533/6346	
Alcohol use			
No	1413/4828	1917/6346	0.2813
Yes	3415/4828	4429/6346	
Diagnosed hypertension			
No	3916/5631	4463/6346	0.3502
Yes	1715/5631	1883/6346	
Self-measured BP			
No	4359/5634	4944/6346	0.4806
Yes	1275/5634	1402/6346	
Advised, self-measured BP			
No	4978/5635	5635/6346	0.4341
Yes	657/5635	711/6346	
Diagnosed hypercholesterolemia			
No	3995/5608	4548/6346	0.6037
Yes	1613/5608	1798/6346	
Gastrointestinal disease			
No	5462/5862	5899/6346	0.6324
Yes	400/5862	447/6346	
Self-reported dietary health			
Excellent	492/5636	594/6346	0.2301
Very good	1209/5636	1412/6346	0.2911
Good	2383/5636	2647/6346	0.5277
Fair	1257/5636	1376/6346	0.4133

Poor	295/5636	317/6346	0.5532
Diagnosed asthma			
No	5303/6340	5309/6346	0.9812
Yes	1037/6340	1037/6346	
Diagnosed anemia			
No	6105/6343	6108/6346	0.9958
Yes	238/6343	238/6346	
Diagnosed psoriasis			
No	5502/5632	6200/6346	0.9780
Yes	130/5632	146/6346	
Diagnosed celiac disease			
No	6312/6341	6317/6346	0.9976
Yes	29/6341	29/6346	
Gluten-free diet			
No	6240/6345	6241/6346	0.9991
Yes	105/6345	105/6346	
Diagnosed arthritis			
No	3726/4951	4763/6346	0.8050
Yes	1225/4951	1583/6346	
Diagnosed gout			
No	4787/4959	6127/6346	0.9598
Yes	172/4959	219/6346	
Diagnosed congestive heart failure			
No	4827/4958	6183/6346	0.8071
Yes	131/4958	163/6346	
Diagnosed coronary heart disease			
No	4795/4950	6153/6346	0.7835
Yes	155/4950	193/6346	
Diagnosed angina			
No	4861/4956	6213/6346	0.5020
Yes	95/4956	133/6346	
Diagnosed heart attack			
No	4805/4955	6153/6346	0.9656
Yes	150/4955	193/6346	
Diagnosed stroke			
No	4810/4956	6140/6346	0.3620
Yes	146/4956	206/6346	
Diagnosed emphysema			
No	4886/4960	6240/6346	0.4520
Yes	74/4960	106/6346	
Diagnosed thyroid disease			
No	4449/4951	5712/6346	0.7941
Yes	502/4951	634/6346	
Diagnosed chronic bronchitis			
No	4681/4956	5995/6346	0.9673
Yes	275/4956	351/6346	
Diagnosed liver disease			
No	4766/4954	6104/6346	0.9593
Yes	188/4954	242/6346	
Diagnosed COPD			
No	4810/4959	6144/6346	0.5872
Yes	149/4959	202/6346	
Diagnosed jaundice			
No	6163/6291	6217/6346	0.9940
Yes	128/6291	129/6346	
Diagnosed cancer			

No	4515/4961	5782/6346	0.8494
Yes	446/4961	564/6346	
Familial heart attack			
No	4325/4842	5654/6346	0.7015
Yes	517/4842	692/6346	
Familial asthma			
No	4770/6214	4875/6346	0.9388
Yes	1444/6214	1471/6346	
Familial DM			
No	3021/4858	3925/6346	0.7165
Yes	1837/4858	2421/6346	
Advised to lose weight			
No	4328/5635	4868/6346	0.9012
Yes	1307/5635	1478/6346	
Advised to exercise			
No	3849/5636	4329/6346	0.9281
Yes	1787/5636	2017/6346	
Advised to reduce salt intake			
No	4450/5628	5021/6346	0.9446
Yes	1178/5628	1325/6346	
Advised to reduce fat intake			
No	4283/5632	4854/6346	0.5707
Yes	1349/5632	1492/6346	
Managing weight			
No	2290/5632	2584/6346	0.9485
Yes	3342/5632	3762/6346	
Increasing exercise			
No	2323/5635	2624/6346	0.8902
Yes	3312/5635	3722/6346	
Reducing salt intake			
No	2949/5634	3341/6346	0.7391
Yes	2685/5634	3005/6346	
Reducing fat intake			
No	2714/5632	3056/6346	0.9716
Yes	2918/5632	3290/6346	
Ever smoking			
No	2941/5021	3750/6346	0.5771
Yes	2080/5021	2596/6346	
Past smoking			
No	4625/5844	5001/6346	0.6498
Yes	1219/5844	1345/6346	
Past smokeless tobacco use			
No	5753/5844	6251/6346	0.7867
Yes	91/5844	95/6346	
Past any tobacco use			
No	3599/4893	4921/6346	**<0.0001**
Yes	1294/4893	1425/6346	
Pulse character			
Regular	6033/6132	6242/6346	0.9144
Irregular	99/6132	104/6346	
Hepatitis A antibody			
Negative	2961/5972	3145/6346	0.9800
Positive	3011/5972	3201/6346	
Hepatitis B core antibody			
Negative	5592/5972	5966/6346	0.3873
Positive	380/5972	380/6346	

Hepatitis B surface antigen			
Negative	5933/5969	6309/6346	0.8847
Positive	36/5969	37/6346	
Hepatitis D (anti-HDV)			
Negative	5952/5969	6327/6346	0.8808
Positive	17/5969	19/6346	
Hepatitis B surface antibody			
Negative	4286/5970	4571/6346	0.7696
Positive	1684/5970	1775/6346	
Hepatitis E IgG (anti-HEV)			
Negative	5700/5972	6061/6346	0.8653
Positive	272/5972	285/6346	
Hepatitis E IgM (anti-HEV)			
Negative	5933/5972	6306/6346	0.8745
Positive	39/5972	40/6346	
Tissue transglutaminase (IgA-TTG)			
negative	5887/5913	6315/6346	0.6915
Positive/weakly positive	26/5913	31/6346	
Education			
<9th grade	938/6341	940/6346	0.9749
9-11 grade	1240/6341	1241/6346	0.9996
High school	1250/6341	1250/6346	0.9825
College/AA degree/college graduate/above	2913/6341	2915/6346	0.9958
Creatine kinase			
No	2891/5867	3138/6346	0.8486
Yes	2976/5867	3208/6346	
HPV vaccination			
No	3854/4636	5355/6346	0.0783
Yes	782/4636	991/6346	
Urinary leakage			
No	3036/4513	4420/6346	**0.0085**
Yes	1477/4513	1926/6346	
Pesticide			
No	5013/5898	5391/6346	0.9460
Yes	885/5898	955/6346	
Bilateral ovariectomy			
No	5654/5911	6065/6346	0.8286
Yes	257/5911	281/6346	
Vigorous activity			
No	3695/6317	3710/6346	0.9718
Yes	2622/6317	2636/6346	
Moderate activity			
No	1851/6317	1853/6346	0.8992
Yes	4466/6317	4493/6346	
Functional limitation			
No	3711/4961	4874/6346	**0.0135**
Yes	1250/4961	1472/6346	
Sleeping trouble			
No	4210/5635	4801/6346	0.2332
Yes	1425/5635	1545/6346	
Secondhand smoking			
No	4172/5924	4468/6346	0.9818
Yes	1752/5924	1878/6346	
Gestational DM			
No	5354/5509	6174/6346	0.7322
Yes	155/5509	172/6346	

Overweight baby at birth (>9lb)			
No	5143/5424	6019/6346	0.9458
Yes	281/5424	327/6346	
Hysterectomy			
No	5435/5925	5821/6346	0.9954
Yes	490/5925	525/6346	
Oral contraception			
No	4352/6067	4531/6346	0.6810
Yes	1715/6067	1815/6346	
Female hormones intake			
No	5485/5931	5864/6346	0.8743
Yes	446/5931	482/6346	

Objective 2

Table 5: Variables included in models for the objective 2 and the rationale/evidence for their inclusion

	Variable	Explanation of the NHANES variable and/or its construction and ascribed codes	Evidence/rationale
1	Age	Age in years of the participant at the time of screening. Individuals 80 and over are top coded at 80 years of age.	See no: 5 of the table 3 in the appendix
2	Income-poverty ratio	Ratio of family income to poverty (range of values 0 to 5)	See no: 15 of the table 3 in the appendix
3	Self-rated oral health	Question from the NHANES questionnaire OHQ845: Overall, how would you rate the health of your teeth and gums? Responses were coded as 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor	A case-control study revealed that individuals with T2DM exhibited poorer oral health than controls (Sandberg et al, 2000). Another study reported that the self-perceived oral health was poorer in children with T1DM than controls (Javed et al, 2009).
4	Number of teeth	Total number of intact teeth on examination; Derived from the NHANES dental examination data (range 0-32)	An association between tooth loss and diabetes was reported in a multicenter, population-based study (Greenblatt et al, 2016). A case-control study found that tooth loss is higher in diabetic patients than nondiabetics and increases more in diabetics as both groups grow older (Ikimi et al, 2017).
5	Time since last dental visit	Question from the NHANES questionnaire OHQ030: About how long has it been since you last visited a dentist? Responses were coded as 1 = 6 months or less, 2 = More than 6 months, but not more than 1 year ago, 3 = More than 1 year, but not more than 2 years ago, 4 = More than 2 years, but not more than 3 years ago, 5 = More than 3 years, but not more than 5 years ago, 6 = More than 5 years ago, 7 = Never have been	Regular receipt of dental care may reduce the need for diabetes-related healthcare (Mosen et al, 2012).
6	Dental floss use frequency	Question from the NHANES questionnaire OHQ870: Aside from brushing your teeth with a toothbrush, in the last seven days, how many days did you use dental floss or any other device to clean between your teeth? Data were collected from individuals of 30 years of age and older	A study by Abduljabbar et al (2017) revealed that periodontal and peri-implant inflammatory parameters were worse among patients with prediabetes and T2DM compared with controls but comparable among patients with prediabetes and T2DM. Dental floss is an interdental cleaning device which helps maintain periodontal health.
7	Mouthwash use frequency	Question from the NHANES questionnaire OHQ875: Aside from brushing your teeth with a toothbrush, in the last seven days, how many days did you use mouthwash or other dental rinse product that you use to treat dental disease or	A clinical trial reported of a significant improvement in periodontal health parameters in the experimental group which used 0.1% chlorhexidine oral rinse

		dental problems? Data were collected from individuals of 30 years of age and older	(Vechis-Bon, 1989).
8	Periodontitis	Data on periodontal pockets, recession, and loss of attachment are available for individuals of 30 years of age and older. Periodontitis was defined as per Eke et al (2012) and initially classified into four groups as healthy, mild, moderate, and severe periodontitis and subsequently dichotomized by aggregating moderate and severe forms to create "periodontitis" category versus "no periodontitis". Healthy and those with mild periodontitis were classified as "no periodontitis". Eke et al (2012) definitions: severe periodontitis = 2 or more interproximal sites with >=6mm AL (not on the same tooth) and 1 or more interproximal sites with >=5mm PD; moderate periodontitis = 2 or more interproximal sites with >=4mm AL (not on the same tooth) or 2 or more interproximal sites with PD>=5mm, also not on the same tooth; mild periodontitis = >=2 interproximal sites with >=3mm AL and >=2 interproximal sites with >=4mm PD (not on the same tooth) or 1 site with >= 5mm Final coding was: 1 = no periodontitis, 2 = periodontitis	See no: 6 and no: 7 above. A bidirectional association is observed between diabetes and periodontal diseases (Mealey et al, 2006). Impaired fasting glucose and/or prediabetes are strongly associated with bleeding on probing, a marker of chronic gingival/periodontal inflammation (Andriankaja & Joshipura, 2014). Periodontal inflammation was found to be associated with an increased risk of pre-diabetes and subsequent incident diabetes (Mustapha, 2014). Higher colonization levels of specific periodontopathic bacteria are observed among prediabetic individuals (Demmer et al, 2015). A metabolomic analysis revealed that experimental periodontitis results in prediabetes (Ilievski et al, 2016). A cross-sectional study revealed that periodontitis is a possible early sign of diabetes mellitus (Teeuw et al, 2017). An experimental study revealed that Prediabetes enhances periodontal inflammation (Huang et al, 2016).
9	Gender	Gender of the participant Responses were coded as 1 = female, 2 = male	See no: 4 of the table 3 in the appendix
10	Race	The variable was dichotomized as 1 = (non-Hispanic) White, 2 = other	See no: 6 of the table 3 in the appendix
11	Citizen	Question from the NHANES questionnaire: "Are you a citizen of the United States?" Responses were coded as 1 = yes, 2 = no.	See no: 9 of the table 3 in the appendix
12	Marital status	The variable was dichotomized as 1 = unmarried/other, 2 = married/ living with partner	See no: 10 of the table 3 in the appendix
13	Ever had periodontal treatment	Question from the NHANES questionnaire OHQ850: Have you ever had treatment for gum disease such as scaling and root planing, sometimes called "deep cleaning"? Data were collected from individuals of 30 years of age and older. Responses were coded as 1 = no, 2 = yes	Periodontal treatment improves the glycemic status (Altamash, 2016)
14	Self-reported tooth mobility	Question from the NHANES questionnaire OHQ855: Have you ever had any teeth become loose on their own, without an injury? Data were collected from individuals of 30 years of age and older. Responses were coded as 1 = no, 2 = yes	Diabetes increases periodontal bone loss (Liu et al, 2006). A positive association was revealed between the duration of diabetes and clinical attachment loss (Firatli et al, 1996). Diabetes is an independent risk factor for clinical attachment loss (Grossi et al, 1994). Attachment

			loss is a key clinical parameter used in periodontal disease assessment and leads to the clinical manifestation of periodontopathic tooth mobility
15	Education	Codes were ascribed as follows: 1 = < 9th grade, 2 = 9-11 grade, 3 = high school, 4 = college/ AA degree/ college graduate/ above	See no:135 of the table 3 in the appendix

References

Abduljabbar, T., Al-Sahaly, F., Al-Kathami, M., Afzal, S. and Vohra, F., 2017. Comparison of periodontal and peri-implant inflammatory parameters among patients with prediabetes, type 2 diabetes mellitus and non-diabetic controls. *Acta Odontologica Scandinavica*, *75*(5), pp.319-324.

Altamash, M., 2016. Periodontal conditions and treatment outcomes for subjects with diabetes mellitus: special emphasis on HbA1c levels and T-cells.

Andriankaja, O.M. and Joshipura, K., 2014. Potential association between prediabetic conditions and gingival and/or periodontal inflammation. *Journal of diabetes investigation*, *5*(1), pp.108-114.

Demmer, R.T., Jacobs Jr, D.R., Singh, R., Zuk, A., Rosenbaum, M., Papapanou, P.N. and Desvarieux, M., 2015. Periodontal bacteria and prediabetes prevalence in ORIGINS: the oral infections, glucose intolerance, and insulin resistance study. *Journal of dental research*, *94*(9_suppl), pp.201S-211S.

Eke, P.I., Dye, B.A., Wei, L., Thornton-Evans, G.O. and Genco, R.J., 2012. Prevalence of periodontitis in adults in the United States: 2009 and 2010. *Journal of dental research*, *91*(10), pp.914-920.

Firatli, E., Yilmaz, O. and Onan, U., 1996. The relationship between clinical attachment loss and the duration of insulin-dependent diabetes mellitus (IDDM) in children and adolescents. *Journal of clinical periodontology*, *23*(4), pp.362-366.

Greenblatt, A.P., Salazar, C.R., Northridge, M.E., Kaplan, R.C., Taylor, G.W., Finlayson, T.L., Qi, Q. and Badner, V., 2016. Association of diabetes with tooth loss in Hispanic/Latino adults: findings from the Hispanic Community Health Study/Study of Latinos. *BMJ Open Diabetes Research and Care*, *4*(1), p.e000211.

Grossi, S.G., Zambon, J.J., Ho, A.W., Koch, G., Dunford, R.G., Machtei, E.E., Norderyd, O.M. and Genco, R.J., 1994. Assessment of risk for periodontal disease. I. Risk indicators for attachment loss. *Journal of periodontology*, *65*(3), pp.260-267.

Huang, Y., Guo, W., Zeng, J., Chen, G., Sun, W., Zhang, X. and Tian, W., 2016. Prediabetes Enhances Periodontal Inflammation Consistent With Activation of Toll-Like Receptor–Mediated Nuclear Factor-κB Pathway in Rats. *Journal of periodontology*, *87*(5), pp.e64-e74.

Ikimi, N.U., Sorunke, M.E., Onigbinde, O.O., Adetoye, J.O. and Amrore, I., 2017. A Study of the Relationship between Diabetes Mellitus and Tooth Loss among Diabetic Patents in Garki General Hospital Garki Abuja, Fct Nigeria. *Dentistry*, *7*(439), pp.2161-1122.

Ilievski, V., Kinchen, J.M., Prabhu, R., Rim, F., Leoni, L., Unterman, T.G. and Watanabe, K., 2016. Experimental periodontitis results in prediabetes and metabolic alterations in brain, liver and heart: global untargeted metabolomic analyses. *Journal of oral biology (Northborough, Mass.)*, *3*(1).

Javed, F., Sundin, U., Altamash, M., Klinge, B. and Engström, P.E., 2009. Self-perceived oral health and salivary proteins in children with type 1 diabetes. *Journal of oral rehabilitation*, *36*(1), pp.39-44.

Liu, R., Bal, H.S., Desta, T., Krothapalli, N., Alyassi, M., Luan, Q. and Graves, D.T., 2006. Diabetes enhances periodontal bone loss through enhanced resorption and diminished bone formation. *Journal of dental research*, *85*(6), pp.510-514.

Mealey, B.L. and Oates, T.W., 2006. Diabetes mellitus and periodontal diseases. *Journal of periodontology*, *77*(8), pp.1289-1303.

Mosen, D.M., Pihlstrom, D.J., Snyder, J.J. and Shuster, E., 2012. Assessing the association between receipt of dental care, diabetes control measures and health care utilization. *The Journal of the American Dental Association*, *143*(1), pp.20-30.

Mustapha, I.Z., 2014. *Periodontal disease and the risk of prediabetes and type 2 diabetes* (Doctoral dissertation).

Sandberg, G.E., Sundberg, H.E., Fjellstrom, C.A. and Wikblad, K.F., 2000. Type 2 diabetes and oral health: a comparison between diabetic and non-diabetic subjects. *Diabetes research and clinical practice*, *50*(1), pp.27-34.

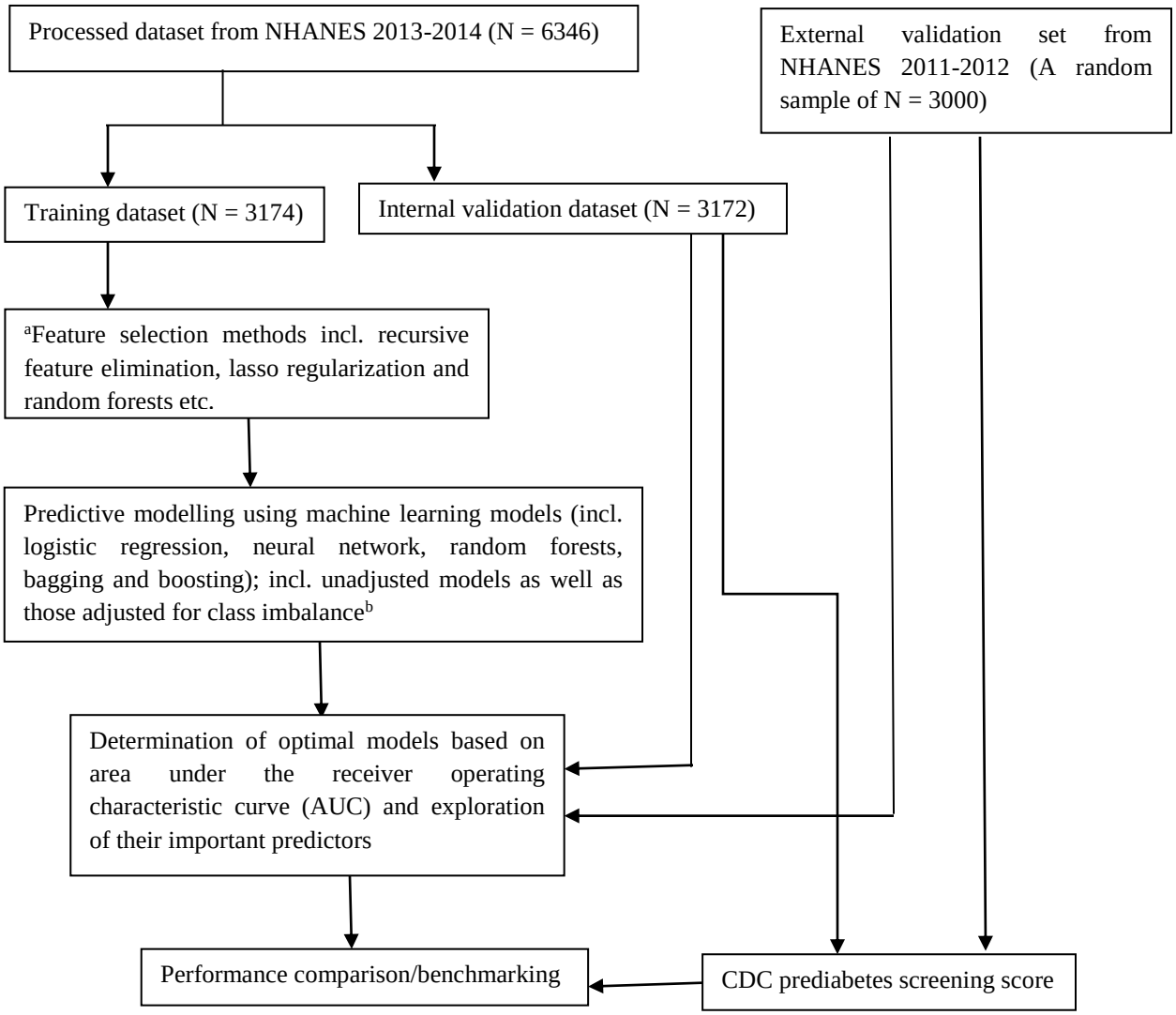
Teeuw, W.J., Kosho, M.X., Poland, D.C., Gerdes, V.E. and Loos, B.G., 2017. Periodontitis as a possible early sign of diabetes mellitus. *BMJ Open Diabetes Research and Care*, *5*(1), p.e000326.

Vechis-Bon, S., 1989. 0.1% chlorhexidine mouthwash for gingival inflammation in diabetic adults: double blind study. *Journal de parodontologie*, *8*(3), pp.299-310.

Table 6: Distribution of variables with missing data in original and multiply-imputed datasets prepared for the objective 2 and post-hoc accuracy analysis of the multiple imputation algorithm

Numeric variables			
Variable	**Original sample mean; SD (N)**	**Imputed sample mean; SD (N)**	**p-value***
Income-poverty ratio	2.71; 1.68 (2920)	2.69; 1.67 (3167)	0.6416
Self-rated oral health[a]	2.98; 1.12 (3165)	2.98; 1.12 (3167)	1.0000
Time since last dental visit[b]	2.54; 1.85 (3166)	2.54; 1.85 (3167)	1.0000
Dental floss/device use frequency[c]	3.46; 2.93 (3166)	3.46; 2.93 (3167)	1.0000
Categorical variables			
Variable	**Original sample proportion**	**Imputed sample proportion**	**p-value***
Citizen of United States			
Yes	2724/3160	2728/3167	0.9410
No	436/3160	439/3167	0.9410
Marital status			
Married/living with partner	2260/3166	2260/3167	0.9842
Widowed/Divorced/Separated/Never married	906/3166	907/3167	
Ever had periodontal treatment			
Yes	756/3159	758/3167	0.9980
No	2403/3159	2409/3167	
Self-reported tooth mobility			
Yes	459/3165	459/3167	0.9917
No	2706/3165	2708/3167	
Education			
<9th grade	223/3166	224/3167	0.9636
9-11 grade	371/3166	371/3167	0.9963
High school	675/3166	675/3167	0.9948
College/AA degree/college graduate or above	1897/3166	1897/3167	0.9877

***Distributions of numeric and categorical variables between original and imputed samples were compared using two-samples t-test and chi-squared test for two proportions, respectively. The p-value of significance was set at 0.05, a-entered as a continuous variable scaled 1-5; see appendix: table 5 (no:3), b-entered as a continuous variable scaled 1-7; see appendix: table 5 (no:5), c-number of days during the past week dental floss was used; see appendix: table 5 (no:6)**



a-excluded in objective 2(predictive modelling using dental parameters) due to low number of covariates; b-adjusted by resampling methods incl. oversampling, under-sampling, random oversampling (ROSE) and synthetic minority oversampling technique (SMOTE)

Figure 1: flow chart illustrating the steps of predictive modelling using general predictors [objective 1] and benchmarking [objective 3]. (Of note, the same methods were applied for predictive modelling using dental covariates [objective 2] with different sample sizes)

POPULAR SCIENCE SUMMARY

Machine learning is a discipline that involves a combined use of high computing power of modern computers and statistical techniques. It has been proven that machine learning applications on large medical databases can provide novel insights into various health issues. This study aimed at applying machine learning techniques on a large public health database, namely, the National Health and Nutrition Examination Survey (NHANES) 2013-2014, to identify multiple factors that may affect the development of prediabetes, which is a common disease across the world. Since it is a reversible condition, if identified early, the progression to diabetes can be prevented, and normal blood glucose levels can be achieved. Nevertheless, timely identification of prediabetes is difficult and current screening tools based on a limited number of traditional risk factors may often fail to identify many prediabetic individuals.

Through a machine learning analytic approach, we identified an array of socio-economic, clinical, biochemical, and dental factors influencing prediabetes, many of which had been reported in previous studies. Interestingly, the study further revealed that several known diabetes risk markers may be potential indicators of prediabetes as well, providing new evidence and directions for future research of the disease. The findings indicate that routinely-collected NHANES data by questionnaires and simple tests such as a person's body measurements, vigorous exercise level, blood lipid level, blood pressure, various blood cell measurements, liver function profile and gum disease can help identify people that are highly likely to develop prediabetes and may complement standard prediabetes risk assessment tools.

However, owing to the limitations of the study design, findings do not confirm that these factors "cause" prediabetes. Further research is warranted to consolidate the findings of the present study which, with a higher level of evidence, may eventually be found useful for clinical diagnosis and community-based screening of prediabetes.