HHS Public Access

Author manuscript

Inform Med Unlocked. Author manuscript; available in PMC 2020 August 28.

Published in final edited form as:

Inform Med Unlocked. 2019 ; 17: . doi:10.1016/j.imu.2019.100254.

Development of non-invasive diabetes risk prediction models as decision support tools designed for application in the dental clinical environment

Harshad Hegde, **Neel Shimpi**, **Aloksagar Panny**, **Ingrid Glurich**, **Pamela Christie**, **Amit Acharya***

Center for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI, USA

Abstract

The objective was to develop a predictive model using medical-dental data from an integrated electronic health record (iEHR) to identify individuals with undiagnosed diabetes mellitus (DM) in dental settings. Retrospective data retrieved from Marshfield Clinic Health System's data-warehouse was pre-processed prior to conducting analysis. A subset was extracted from the preprocessed dataset for external evaluation ($N_{validation}$) of derived predictive models. Further, subsets of 30%–70%, 40%–60% and 50%–50% case-to-control ratios were created for training/testing. Feature selection was performed on all datasets. Four machine learning (ML) classifiers were evaluated: logistic regression (LR), multilayer perceptron (MLP), support vector machines (SVM) and random forests (RF). Model performance was evaluated on $N_{validation}$. We retrieved a total of 5319 cases and 36,224 controls. From the initial 116 medical and dental features, 107 were used after performing feature selection. RF applied to the 50%–50% case-control ratio outperformed other predictive models over $N_{validation}$ achieving a total accuracy (94.14%), sensitivity (0.941), specificity (0.943), F-measure (0.941), Mathews-correlation-coefficient (0.885) and area under the receiver operating curve (0.972). Future directions include incorporation of this predictive model into iEHR as a clinical decision support tool to screen and detect patients at risk for DM triggering follow-ups and referrals for integrated care delivery between dentists and physicians.

Keywords

Dental informatics; Decision-support systems; Electronic health records; Evidence-based practice; Machine leaning; Modeling healthcare services

*Corresponding author. Marshfield Clinic Research Institute, Research Scientist, Center for Oral and Systemic Health, Marshfield Clinic Research Institute, 1000, North Oak Avenue, Marshfield, WI, 54449. USA. acharyaa@marshfieldresearch.org (A. Acharya).

Declaration of competing interest

The authors do not have any conflict of interest.

1. Introduction

Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by development of abnormal regulation of blood glucose levels [1]. The disease progresses gradually through a pre-diabetic phase associated with sub-clinical levels of dysglycemia and often goes undetected during initial years due to its asymptomatic nature [1]. Disease progression is associated with incremental increases in magnitude of micro and macro vascular complications [2]. Hence early detection of undiagnosed DM is crucial for prevention of associated complications, progression delay and better overall management of morbidity [3]. DM is a global pandemic with 415 million individuals affected in 2015, with prevalence projected at 642 million by 2040 [4]. In United States, an estimated 30 million individuals were diagnosed with DM in 2017 while 7.2 million of individuals remained undiagnosed [5].

A mounting evidence base has demonstrated bidirectional association between periodontal disease (PD) and DM [6-8]. PD has been recognized as an early complication of DM [9-11]. Population-based screening to identify individuals at risk for dysglycemia may support prevention and intervention. However, biological screening is currently not supported as a standard of care in the dental setting following evaluation of the evidence base by the US Preventive Services Task Force and United Kingdom's National Institute for Health Research in 2008 and 2013 respectively [12]. In the absence of biological screening for dysglycemia in the dental setting, non-invasive screening of existent medical and dental data in the electronic health record (EHR) with capacity to identify patients potentially at risk for DM could be really beneficial [12]. These patients may require referral for further evaluation that may contribute to early diagnosis and an opportunity to modify patient risk for progression and onset of complications [13]. The value of using EHR data to detect DM at the point of care (POC) for identification of high-risk individuals has been demonstrated in a study by Sohler et al. (2016) [14]. Moreover a recent systematic review of 10 field trials that evaluated biological screening for DM in the dental setting to project DM prevalence among patients [15]. Their review reported detection of glycemic measures in the diabetic range at POC among 1.3% and 14% patients across the studies while rates of dysglycemia in the prediabetic range varied widely, ranging from 19% to as high as 90% [15]. Screening for conditions like DM in a dental setting is an important component of disease identification/ prevention that enables integrated care delivery across disciplines [16].

The percentage of dental visits of adults aged 18 and above substantially increased from approximately 40% in the year 2000 to nearly 60% in 2015 [17,18]. Increase in number of patients seeking dental care provides a unique opportunity to screen individuals for high risk of DM at the POC [19]. Recent studies have showcased the willingness of dental providers to screen and monitor patients for risk of systemic conditions including DM, in the dental setting, to contribute to holistic improvement of health outcomes [19-21]. However, lack of knowledge surrounding DM among dental providers, time constraints and provider perceptions surrounding care for patients with DM were some of the historically-identified barriers to integrated care delivery [20,21]. Creating capacity to identify risk factors and their relative contribution to increased likelihood for developing DM can create opportunities for establishing cost-effective interventions. Applying clinical decision support

tools that conduct continuous monitoring for DM risk by screening available data in the health records creates opportunity for more timely and appropriate intervention in the context of integrated healthcare delivery. However, deciphering complex relationships and interactions among multiple risk factors is challenging to compute. This requires creation of computerized models utilizing effective approaches, including application of artificial intelligence. The objective of our study was to engage machine learning (ML) approaches to develop prediction tools which can be implemented at the POC in a dental setting to identify patients at a risk for DM. Implementing such a clinical decision support tool at the (POC) will aid the dental providers in identification of individuals with high risk of DM and directly inform care delivery. Based on our understanding, this is one of the first studies utilizing medical and dental data to develop predictive tool for identifying the risk of DM in dental settings.

2. Methods

2.1. Study setting and population

Marshfield Clinic Health System (MCHS) is one of the largest, comprehensive, medical-dental health systems in the United States with a service area that spans care delivery throughout central, northern and western Wisconsin [22]. Care delivery across its networked medical and dental clinics is supported by a robust integrated medical-dental electronic health record (iEHR) environment. This multi-specialty group practice employs nearly 700 physicians and 40 dentists and approximately 7600 employees who support care delivery. Family Health Center of Marshfield (FHC), a federally qualified health center, partners with MCHS to provide care to more than 59,000 unique dental patients annually through 10 regional dental clinics spanning an extensive service area across Wisconsin which closely aligns with MCHS service area [23,24]. The majority of the populations residing within the largely rural MCHS service area are White/Caucasian. The study was reviewed and approved by using expedited review by the Institutional Review board (IRB) of Marshfield Clinic Research Institute.

2.2. Data retrieval

Retrospective data spanning a 39-year temporal period from 1979 to 2018 were mined from MCHS's enterprise research data warehouse (EDW). A comprehensive list of potential candidate data features were first identified and cataloged by systematic review of previously-published diabetes/dysglycemia risk assessment models. using multivariate regression across other populations [13]. Features representing candidate risk factors retained in historic models and with available data in our iEHR were selected for further modeling to evaluate their relevance and validity relative to our patient population. Notably, if multiple measures of a feature were made, each measure was initially tallied as a separate feature (e.g., PPD or BOP measures made at 6 sites per tooth). Multiple measures of a feature were finally reduced to a single representation for that feature (e.g. we only considered the max PPD value out of the 6 sites). The final dataset included both medical and dental features. Predicting DM risk was treated as a classification problem and outcome was represented by two categories (binary outcome): 'high risk (cases)' and 'low risk (controls)'. Cases were defined as all the patients who were diagnosed with DM identified

by ICD-9/10 codes by the physician practices while controls were defined as patients who were non-diabetics lacking such coding in the EHR.

The following inclusion/exclusion criteria were applied to select individuals for the study.

- Only patients with both medical and dental visits were included in the study.
- All data one year prior to the date of DM diagnosis was collected for cases. For controls all data was collected in a one year time frame prior to the last dental visit.
- The dental data associated with the third molars were excluded from analysis.
- All individuals between 21yrs. of age and 89 yrs of age were included in the study.

2.3. Data preprocessing

2.3.1. Data deletion—The percentage of missing data for each feature and percentage of missing data for each patient record was calculated. Any feature with more than fifty percent of the data missing and any patient record with more thirty percent of the data missing were excluded from the analysis. Table 1 illustrates the list of all data features included in the prediction model development along with baseline characteristics. Table 2 illustrates the deleted data features from the datasets due to a high proportion of missing data (see Table 3).

2.3.2. Redefining feature values—We subset several features into predefined categories [26-28] (e.g. feature #2: BMI in Table 1 were categorized according to study described by Bhaskaran et al. [25] as *underweight (<18.5); normal (*18.5–24.99), *overweight* (25.0–29.99) or *obese(>30*).

For the feature: “periodontal pocket depth” (PPD) we obtained the maximum PPD among the six probing surfaces of the tooth to accurately define the extent of periodontal tissue destruction. The maximum PPD or the worst depth among the six probing sites for each tooth was recorded to classify the severity and extent of a patients periodontal disease along with other factors. The latest classification of periodontal disease was adopted at the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases [29]. The feature “bleeding on probing” (BOP) was recorded as a boolean value that evaluated whether bleeding was present or absent on a tooth’s surface. Moreover this data point was aligned to the corresponding tooth surface that showed the deepest PPD. This exercise resulted in a single value for both “PPD” and “BOP” for each tooth.

2.4. Label encoding and feature scaling [30,31]

All categorical features (string labels) in the preprocessed dataset were transformed to numerical values by importing the data as a Data Frame in R software (R version 3.4.3. R Foundation for Statistical Computing, Vienna, Austria.) [32]. Further, all the features were scaled to a range of [0, 1] in order to prevent any feature being weighted more than others due to a larger range. This normalization was performed using the following function [31]:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Where x_{norm} is the normalized value, x is the original value, max(x) is the upper bound, and min(x) is the lower bound values for the features. All feature scaling was performed using R programming language [32].

2.5. Dataset creation

A validation set ($N_{validation}$) was first separated out from the preprocessed dataset, which constituted of 10% of the total number of cases and equal number of controls. The remaining 90% of training/testing set ($N_{train/testing}$) was further divided randomly into three subsets that included cases and controls combined in ratios of 30:70; 40:60 and 50:50, respectively (Fig. 1). The number of cases were kept consistent (n =4757) and the number of controls were adjusted accordingly. The controls were randomly sampled without replacement for creating these datasets. SAS®(Base SAS 9.4 SAS Institute Inc., Cary, NC) analytical software was used for data preprocessing and datasets creation [33].

2.6. Data imputation

We initially evaluated two data imputation methods: Multiple Imputation by Chained Equations (MICE) [34,35] and Probabilistic Principal Component Analysis (PPCA) [36], to address the missing data values in the $N_{train/testing}$. This was done by selecting a subset of data from $N_{train/testing}$ without any missing data points for all the features. Thirty percent "missingness" was imposed into this subset using 'missing completely at random' (MCAR) mechanism to mimic the missing data pattern in preprocessed dataset. The percentage of values that were correctly imputed was higher using PPCA than MICE; hence PPCA method was employed to impute missing values in $N_{train/testing}$. Data imputation was not carried out in $N_{validation}$ to replicate the real-world scenario for evaluating the classifier performance.

2.7. Feature selection

Feature selection was conducted on all 3 datasets using WEKA® [37]. We performed the feature selection using information gain with ranker search method. This method evaluates the importance of each feature by measuring the information gain with respect to the class. using the following formula:

$$\text{InfoGain(Class, Feature)} = \text{H(Class)} - \text{H(Class} \mid \text{Feature)}. \quad (2)$$

Where H, stands for entropy, which is defined as:

$$H = -\sum (Probability_{class} * \log_2(Probability_{class})) \quad (3)$$

2.8. Model training and validation

Based on previous studies [38] four supervised ML algorithms: Multilayer Perception (MLP), Random Forests (RF), Support Vector Machine (SVM) and Logistic Regression (LR) were used to create models for predicting DM risk. The detailed description of these classifiers with mathematical equations can be found in other papers [39,40]. These classifiers were trained and tested on the three subsets (30:70; 40:60 and 50:50) using 10-fold cross validation. Each fold derived a model whose performance was evaluated over $N_{validation}$ as shown in Fig. 1 and the results of the best performing fold for each corresponding classifier were reported.

WEKA® an open source software tool was used for building these models [37]. All classifiers were implemented using default hyperparameters. Performance metrics to test model performance included evaluation of the total accuracy, sensitivity, specificity, precision, F-measure, Mathews-correlation-coefficient (MCC), false positive rate (FPR), false negative rate (FNR), Negative predictive value (NPV), Positive predictive value (PPV) and area under Receiver Operating Characteristic (ROC) curve (AUC). A paired two-tailed *t*-test was performed over AUC using WEKA Experimenter. The statistical significance was set at $\alpha = 0.05$. To indicate the strength of the difference between cases and controls, the observed p-value was used to determine the association (χ^2-test).

3. Results

Systematic review identified 69 articles published between 1/1980 and 5/2018 that undertook diabetes risk prediction modeling. These studies examined contribution of a total of 201 candidate medical, dental, demographic, environmental and behavioral candidate features. Of these, 95 features shown in Fig. 2 were variably retained in predictive models created across diverse population cohorts. Among 19 dental features, three variables including PPD, missing teeth and self-reported oral health status were retained in some final models as predictors. Availability of these features was explored in our clinical databases for inclusion in model development and available features were retrieved (see Fig. 3).

Data retrieved were abstracted from a total of 41,543 subjects (5319 cases and 36,224 controls) and 124 features which included demographic (n = 17), medical/environmental/behavioral (n = 15) and oral health (n = 92) features, where each measure of a variable was initially abstracted as a discrete feature (e.g. BOP and PPD measured at 6 sites per tooth). Preprocessing of data resulted in deletion of features based on a high proportion of the missing data, or redefinition via Boolean representation (teeth present or absent) or derivation of a representative value across a series of iterative measures of a feature (e. g. PPD). The processed dataset consisted of 40,519 patients (5286 cases and 35,233 controls) and 116 features (18 demographic, 12 medical and 86 dental features) (see Table 1). After performing feature selection on all 3 datasets, race and ethnicity were excluded due to low information gains of $\sim 10^{-3}$ and $\sim 3*10^{-4}$ respectively thus bringing the feature count down from 116 to 107. The most significant feature in terms of information gain for all 3 datasets was 'Number of dental visits' (~0.6). $N_{validation}$ consisted of a total of 529 cases (10% of the total cases (5,286)) and an equal number of controls totaling to 1058 observations. After

separation of $N_{validation}$, the resulting $N_{train/testing}$ consisted of 39,461 subjects with 4757 cases and 34,704 controls.

A total of 12 prediction models (3 datasets 'times' 4 ML classification methods) were developed. RF classifier trained on a dataset with a class distribution of 50-50 (case-control) was the best performing model.

Table 4 shows the confusion matrix of RF with 50-50 (case: control) which was the best performing model. Number of dental visits and PPD were the top two features in terms of information gain. The correctly and incorrectly classified instances using RF were 996 and 62 respectively (see Table 5).

We also assessed the best performing models for each of the remaining classifiers. MLP classifier performed better on the class distribution of 40–60 (case-control) and demonstrated a total accuracy of 82.51% and ROC (AUC) of 0.9. LR classifier was efficient with the class distribution of 40–60 (case-control) and demonstrated a total accuracy of 86.29% and ROC (AUC) of 0.935. Similarly, SVM classifier performed better over the class distribution of 40–60 (case-control), achieving a total accuracy of 85.35% and ROC (AUC) of 0.806.

4. Discussion

DM diagnosis is dependent on demonstration of two measures of elevated plasma glucose. These include either a) fasting plasma glucose (FPG) or b) 2-h plasma glucose (2-h PG) value following administration of a 75-g oral glucose tolerance test (OGTT), or based on Hemoglobin A1C (HbA1C) criteria [1], with diagnoses rendered by physicians in a clinical setting. Although dentists do not diagnose DM, recognition of undiagnosed DM is highly relevant to provision of appropriate oral and dental care to a patient with this condition. With mounting evidence supporting association between DM and PD, development of predictive modeling tools such as the tool defined in this study will add value to dental professionals in identifying undiagnosed patients potentially at high risk for DM [38,41]. The bidirectional nature of DM and PD and the importance of managing PD with the knowledge that the individual may be at risk for DM may be crucial at the POC. Identification of potentially at risk individuals in the dental setting supported by triage to clinicians for further testing and confirmation of DM or prediabetic status in an earlier timeframe which benefits patient health.

The objective of our study was to develop a prediction model to noninvasively screen EHR data of patients with no existing diagnosis for dysglycemia seen in the dental setting to identify presence of relative risk of DM. Predictive modeling using ML algorithms has been previously applied in a number of studies for risk determination of various health conditions [42,43]. A systematic review suggested that the ML algorithms applied in our study aligned with those applied in similar risk prediction modeling studies undertaken by other researchers [38]. We focused on increasing predictive accuracy by modeling a combination of medical and dental risk factor candidates to develop a model to predict DM risk for

patients in our clinical population. Among the classifiers used in this study, RF yielded the best AUC which was statistically significant as compared to the rest ($p < 0.05$).

Acharya et al. previously developed a model using multivariate logistic regression as a tool for screening diabetes risk [13]. The study reported an AUC (ROC), sensitivity and specificity of 0.71, 0.70 and 0.62 respectively. Application of ML and expansion of the candidate predictive features in the current study resulted in improved performance of the models developed that achieved AUC (ROC), sensitivity and specificity of 0.972, 0.941 and 0.915 respectively. Improved performance noted in the current study is likely attributable to the volume of data used to train the models: i.e. 107 features modeled in data set of 39,461 individuals vs. 10 features modeled in 4560 individuals in the study by Acharya et al. [13].

Similar studies were done in the past for developing nomograms to predict the risk of DM based on laboratory, semi-laboratory and nonlaboratory data [44,45]. Wong et al. and Li et al. used logistic regression to develop their models. Wong et al. reported an AUC of 0.709 for non-laboratory-based risk algorithm while 0.711 for laboratory-based risk algorithms respectively. Similarly, Li et al. reported an AUC of 0.868 and 0.763 for the semi-lab model and non-lab model respectively. Both studies used 25% of their data for external evaluation. In this study, we used only 10% of cases and equal number of controls (n = 529 each) as part of the external evaluation set in the interest of using most of the data for training and testing so that the feature space was adequately represented.

Lalla et al. built a model to predict undiagnosed pre-diabetes and DM using LR applied to dental features in a cohort of subjects whose HbA1C results at (POC) classified them into dysglycemic or normoglycemic at a ratio of 36:64 [10]. The authors reported an AUC of 0.65 for their best performing model, which included only dental features. When an optimal cut-offs were defined including 26% of teeth with deep PPD or ≥4 missing teeth the sensitivity of the model was 0.73.

Li et al. used the data from NHANES III survey to construct predictive models for DM risk using classification and regression tree (CART) method [46]. They used a wider range of risk factors and also tested other dental parameters including ‘sum of decayed, missing and filled (DMFTs) tooth surfaces’, ‘time since last dental visit’, ‘self-reported oral health status’ in addition to ‘presence or absence of periodontitis’. By contrast, the present study did not include DMFTs, ‘time since last dental visit’ and ‘self-reported oral health status’. However, we considered the missing teeth data for all 28 teeth as a boolean value ‘tooth present or absent’. Furthermore, we limited consideration to only the total number of dental visits within one year temporal window. Li et al. used 55 features (N = 15,090) and their model had a sensitivity, specificity and AUC of 0.824, 0.528 and 0.72 respectively.

Borrell et al. also used data from NHANES III to develop a LR model to identify features that increased the probability of identifying a dental patient with an undiagnosed DM status [47]. These investigators modeled the following features: presence or absence of periodontitis, self-reported family history of DM, hypertension and cholesterol levels. Family history of DM was one of the features used in our study as well. Periodontal assessments tested in their models included: clinical attachment loss (CAL) and PPD [47].

Our model also incorporated PPD data. However, we did not have CAL measures available for the study.

5. Limitations

Since some features were excluded due to missing data (Table 2) whose proportion was greater than 50%, hence imputation techniques were limited to those shown in Table 1. Our dataset consisted of a predominantly White, non-Hispanic population and hence lacked the racial diversity found in the data used by Borell et al. which consisted of a mixed African-American, Mexican-American (Hispanic) and White racial/ethnic representation. Feature selection on our dataset revealed that race and ethnicity had very low information gain which could be associated to the lack of diversity. The model developed in this study was evaluated with an internal dataset and plans include evaluation of our model using data from external organizations. Although our models were restricted to the data procured from a single health system, the service area of MCHS covers a vast geographic area including central, northern and western Wisconsin. The features used in the present study were captured in a clinical setting and not specifically for research purposes. This resulted in the exclusion of some features modeled in other studies because these data were not systematically available in our clinical setting [46]. For example, Li et al. incorporated waist-circumference for testing in their models, which other studies have found to be a highly predictive feature for dysglycemia in some populations. Moreover, some features that were captured mainly in clinical notes could not be readily modeled due to challenges involved in abstracting them. Abstraction could be facilitated by application of natural language processing (NLP) in order to convert clinical notes into structured features for our future studies [48]. Examples of such features include tooth brushing or flossing frequency, which are recorded in the free text fields but not currently captured as structured data elements. Another limitation that we encountered was missing dental visits for patients who carried a DM diagnosis within the last year of the date of diagnosis. Such potentially informative patients did not meet inclusion criteria due to insufficient data within the temporal frame specified for inclusion in the study.

6. Conclusions

We developed a model that identifies undiagnosed patients at risk for DM based on their medical and dental data. Future direction includes embedding this algorithm in an application programmable interface (API) for incorporation into the EHR to provide alerts to dental providers. Moreover, application of this process could be employed to develop additional decision support tools for other systemic diseases with oral health associations such as cardiovascular and cognitive disorders. All features used in modeling performed to date represent phenotypic and demographic data. Additionally relevant genetic data could be modeled in the future to determine whether such data would increase predictive capacity of current clinical phenotypic models as such, data become available in the iEHR.

Acknowledgments

Funding sources

This work was supported by funds from Delta Dental of Wisconsin; Marshfield Clinic Research Institute, Family Health Center of Marshfield, Inc. and partial funding by grant UL1TR000427 from Clinical and Translational Science Award (CTSA) program of the National Center for Advancing Translational Sciences, National Institutes of Health (NIH). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1]. American Diabetes Association AD. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. 1 1 [cited 2018 Sep 19] Diabetes Care [Internet] 2018;41(Supplement 1). S13–27. Available from: http://care.diabetesjournals.org/lookup/doi/10.2337/dc18-S002.

[2]. Bergman M Pathophysiology of prediabetes and treatment implications for the prevention of type 2 diabetes mellitus. 6 7 [cited 2018 Jul 20] Endocrine [Internet] 2013;43(3). 504–13. Available from: http://link.springer.com/10.1007/s12020-012-9830-9.

[3]. American Diabetes Association. 5. Prevention or delay of type 2 diabetes. 1 1 [cited 2018 Jul 20] Diabetes Care [Internet] 2015;38(Supplement 1). S31–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25537704.

[4]. Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. 6 1 [cited 2018 Sep 17] Diabetes Res Clin Pract [Internet] 2017;128:40–50. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168822717303753.

[5]. National diabetes statistics report. 2017 [Internet]. 2017 [cited 2017 Sep 27]. Available from: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

[6]. Lamster IB, Cheng B, Burkett S, Lalla E. Periodontal findings in individuals with newly identified pre-diabetes or diabetes mellitus [cited 2017 Oct 31] J Clin Periodontol [Internet] 2014 11;41(11). 1055–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25195497.

[7]. Wang T-F, Jen I-A, Chou C, Lei Y-P. Effects of periodontal therapy on metabolic control in patients with type 2 diabetes mellitus and periodontal disease: a meta-analysis [cited 2018 Sep 19] Medicine (Baltimore) [Internet] 2014 12;93(28). e292 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25526470.

[8]. Corbella S, Francetti L, Taschieri S, De Siena F, Fabbro M Del. Effect of periodontal treatment on glycemic control of patients with diabetes: a systematic review and meta-analysis. 9 13 [cited 2018 Sep 19] J Diabetes Investig [Internet] 2013;4(5). 502–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24843701.

[9]. Panny A, Krueger K, Acharya A. Achieving the ‘True’ triple aim in healthcare [cited 2019 Jul 10]. In: Amit Acharya, Valerie Powell, Torres-Urquidy Miguel H, Posteraro Robert Hugh, Paul Thyvalikakath Thankam, editors. Integration of medical and dental care and patient data [Internet]. Second. Switzerland: Springer International Publishing; 2019 p. 11–32. Available from:. http://link.springer.com/10.1007/978-3-319-98298-4_2.

[10]. Lalla E, Kunzel C, Burkett S, Cheng B, Lamster IB. Identification of unrecognized diabetes and pre-diabetes in a dental setting [cited 2017 Sep 27] J Dent Res [Internet] 2011 7 29;90(7). 855–60. Available from: http://journals.sagepub.com/doi/10.1177/0022034511407069.

[11]. Lamster IB, Kunzel C, Lalla E. Diabetes mellitus and oral health care: time for the next step. Mar [cited 2017 Sep 27] J Am Dent Assoc [Internet] 2012;143(3). 208–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22383195.

[12]. Glurich I, Nycz G, Acharya A. Status update on translation of integrated primary dental-medical care delivery for management of diabetic patients [cited 2018 Sep 19] Clin Med Res [Internet] 2017 6 1;15(1–2):21–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28373288.

[13]. Acharya A, Cheng B, Koralkar R, Olson B, Lamster IB, Kunzel C, et al. Screening for diabetes risk using integrated dental and medical electronic health record data [cited 2018 Nov 1] JDR Clin Transl Res [Internet] 2018 4 26;3(2). 188–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29568804.

[14]. Sohler N, Matti-Orozco B, Young E, Li X, Gregg EW, Ali MK, et al. Opportunistic screening for diabetes and prediabetes using hemoglobin A1C in an urban primary care setting [cited 2018 Jul

20] Endocr Pract [Internet] 2016 2;22(2). 143–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26484404.

[15]. Glurich I, Bartkowiak B, Berg RL, Acharya A. Screening for dysglycaemia in dental primary care practice settings: systematic review of the evidence. 12 1 [cited 2018 Nov 28] Int Dent J [Internet] 2018;68(6):369–77. 10.1111/idj.12405.

[16]. Genco RJ, Schifferle RE, Dunford RG, Falkner KL, Hsu WC, Balukjian J. Screening for diabetes mellitus in dental practices [cited 2017 Oct 31] J Am Dent Assoc [Internet] 2014 1;145(1):57–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24379330.

[17]. Franco et al. National center for health Statistics.Health. United States. 2016 Available from: https://www.cdc.gov/nchs/data/hus/hus16.pdf#078.

[18]. Nasseh K, Vujicic M. Dental care utilization steady among working-age adults and children, up slightly among the elderly [cited 2018 Jul 20]; Available from: http://www.ada.org/~/media/ADA/ScienceandResearch/HPI/Files/HPIBrief_1016_1.pdf; 2016.

[19]. Shimpi N, Schroeder D, Ph C, Glurich I, Acharya. Assessment of dental providers' knowledge, behavior and attitude towards incorporating chairside screening for medical conditions: a pilot study [cited 2018 Sep 25]; Available from: www.annexpublishers.com.

[20]. Shimpi N, Bharatkumar A, Jethwani M, Chyou P-H, Glurich I, Blamer J, et al. Knowledgeability, attitude and behavior of primary care providers towards oral cancer: a pilot study [cited 2018 Aug 2] J Cancer Educ [Internet] 2018 4 23;33 (2). 359–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27448614.

[21]. Glurich I, Schwei KM, Lindberg S, Shimpi N, Acharya A. Integrating medical-dental care for diabetic patients: qualitative assessment of provider perspectives. 7 26 [cited 2018 Aug 2] Health Promot Pract [Internet] 2018;19(4). 531–41. Available from: http://journals.sagepub.com/doi/10.1177/1524839917737752.

[22]. Shimpi N, Glurich I, Acharya A. Integrated care case study: Marshfield clinic health system [cited 2019 Jul 10], 315–26. Available from: http://link.springer.com/10.1007/978-3-319-98298-4_17; 2019.

[23]. Shimpi N, Ye Z, Koralkar R, Glurich I, Acharya A. Need for diagnostic-centric care in dentistry [cited 2018 Sep 25] J Am Dent Assoc [Internet] 2018 2;149(2). 122–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29389335.

[24]. Acharya A Marshfield clinic health system: integrated care case study. 3 [cited 2018 Sep 25] J Calif Dent Assoc [Internet] 2016;44(3). 177–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27044239.

[25]. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. 8 [cited 2018 Nov 14] Lancet [Internet] 2014;384(9945). 755–65. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673614608928.

[26]. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. New ACC/AHA high blood pressure guidelines lower definition of hypertension. 5 [cited 2018 Nov 14] J Am Coll Cardiol [Internet] 2018;71(19). e127–248. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0735109717415191.

[27]. Berry J, Murrell D. High HDL levels: recommendations, balance, and tips [Internet]. Medical News Today; 2017 [cited 2018 Nov 14]. Available from: https://www.medicalnewstoday.com/articles/319275.php.

[28]. Naushad H, Marion S. Leukocyte count (WBC): reference range, interpretation, collection and panels [Internet] MedScape 2015 [cited 2018 Nov 14]. Available from: https://emedicine.medscape.com/article/2054452-overview#a2.

[29]. Jepsen S, Caton JG, Albandar JM, Bissada NF, Bouchard P, Cortellini P, et al. Periodontal manifestations of systemic diseases and developmental and acquired conditions: consensus report of workgroup 3 of the 2017 world Workshop on the classification of periodontal and peri-implant diseases and conditions [cited 2019 Oct 10] J Clin Periodontol [Internet] 2018 6;45 10.1111/jcpe.12951.

[30]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python [cited 2018 Aug 2] J Mach Learn Res [Internet] 2011; 12 (Oct): 2825–30. Available from: http://www.jmlr.org/papers/v12/pedregosa11a.html.

[31]. Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. 4 1 [cited 2018 Aug 2] Pattern Recognit Lett [Internet] 2001;22(5). 563–82. Available from: https://www.sciencedirect.com/science/article/pii/S0167865500001124.

[32]. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018 Available from: https://www.r-project.org/.

[33]. Base SAS ® 9.4. Procedures guide statistical procedures [Internet]. Cary, NC, USA second ed. 2013 [cited 2019 Jul 11]. Available from: https://support.sas.com/documentation/cdl/en/procstat/66703/PDF/default/procstat.pdf.

[34]. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. 3 [cited 2018 Aug 20] Int J Methods Psychiatr Res [Internet] 2011;20(1). 40–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21499542.

[35]. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. 3 15 [cited 2018 Aug 20] Am J Epidemiol [Internet] 2014;179(6). 764–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24589914.

[36]. Tipping ME, Christopher MB. Probabilistic principal component analysis [cited 2018 Aug 20] J R Stat [Internet] 1999;61(3):611–22. Available from: http://www2.mta.ac.il/~gideon/courses/machine_learning_seminar/papers/ppca.pdf.

[37]. Witten IH, Ian H, Frank E, Hall MA, Mark A, Pal CJ. Data mining: practical machine learning tools and techniques. Fourth. Morgan Kaufmann; 2016 621 pp.

[38]. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. 5 30 [cited 2018 Aug 3] J Med Internet Res [Internet] 2018;20(5). e10775 Available from: http://www.jmir.org/2018/5/e10775/.

[39]. Breiman L Random forests [cited 2019 Jul 19] Mach Learn [Internet] 2001;45(1): 5–32. Available from: http://link.springer.com/10.1023/A:1010933404324.

[40]. Vapnik VN. The nature of statistical learning theory [Internet]. New York, NY: Springer New York; 2000 [cited 2019 Jul 19]. Available from: http://link.springer.com/10.1007/978-1-4757-3264-1.

[41]. Lee J-H, Kim D-H, Jeong S-N, Choi S-H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. 10 1 [cited 2018 Nov 28] J Dent [Internet] 2018;77 106–11. Available from: https://www.sciencedirect.com/science/article/pii/S0300571218302252.

[42]. Shimpi N, McRoy S, Zhao H, Wu M, Acharya A. Development of a periodontitis risk assessment model for primary care providers in an interdisciplinary setting [cited 2019 Jul 10];Preprint(Preprint) Technol Heal Care [Internet] 2019 7 1:1–12. Available from: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/THC-191642.

[43]. Naushad SM, Hussain T, Indumathi B, Samreen K, Alrokayan SA, Kutala VK. Machine learning algorithm-based risk prediction model of coronary artery disease. Mol Biol Rep [Internet] 2018 7 11 [cited 2018 Aug 3]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/29995270.

[44]. Wong CKH, Siu S-C, Wan EYF, Jiao F-F, Yu EYT, Fung CSC, et al. Simple non-laboratory- and laboratory-based risk assessment algorithms and nomogram for detecting undiagnosed diabetes mellitus. 5 1 [cited 2019 Oct 10] J Diabetes [Internet] 2016;8(3):414–21. 10.1111/1753-0407.12310.

[45]. Li W, Xie B, Qiu S, Huang X, Chen J, Wang X, et al. Non-lab and semi-lab algorithms for screening undiagnosed diabetes: a cross-sectional study. 9 [cited 2019 Oct 10] EBioMedicine [Internet] 2018;35 307–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30115607.

[46]. Li S, Williams PL, Douglass CW. Development of a clinical guideline to predict undiagnosed diabetes in dental patients [cited 2017 Sep 27] J Am Dent Assoc [Internet] 2011 1;142(1):28–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21193764.

[47]. Borrell LN, Kunzel C, Lamster I, Lalla E. Diabetes in the dental office: using NHANES III to estimate the probability of undiagnosed disease [cited 2017 Sep 27] J Periodontal Res [Internet] 2007 12 1;42(6):559–65. 10.1111/j.1600-0765.2007.00983.x.

[48]. Hegde H, Shimpi N, Glurich I, Acharya A. Tobacco use status from clinical notes using Natural Language Processing and rule based algorithm. 6 29 [cited 2019 Jul 8] Technol Heal Care [Internet] 2018;26(3). 445–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29614708.

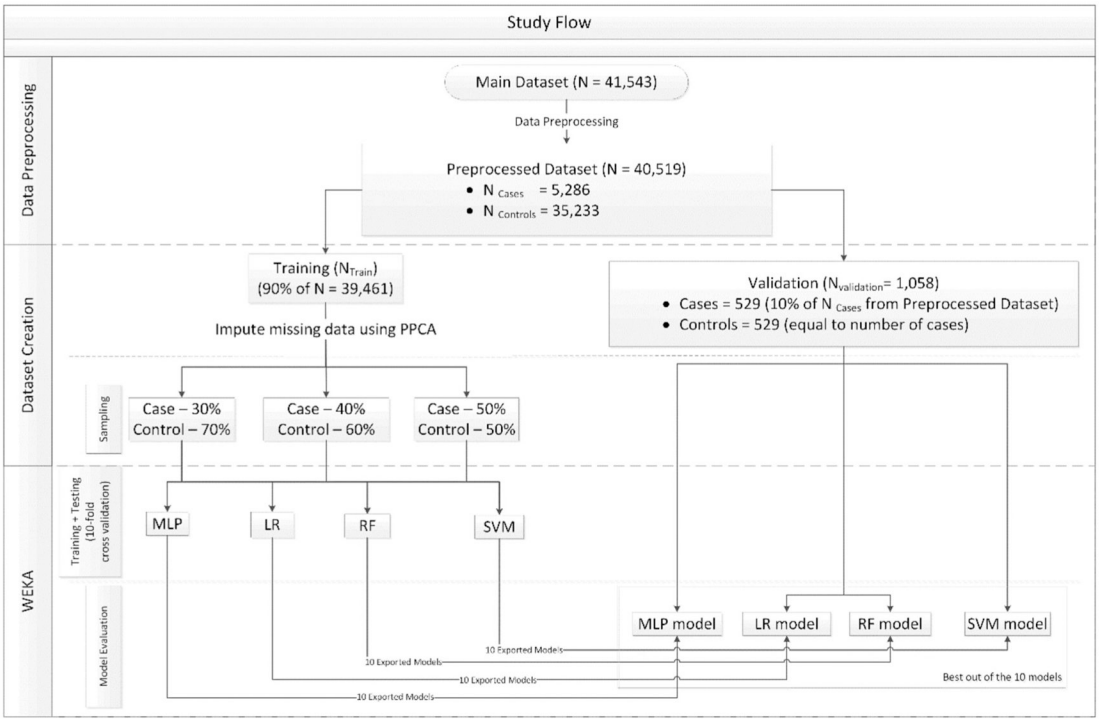
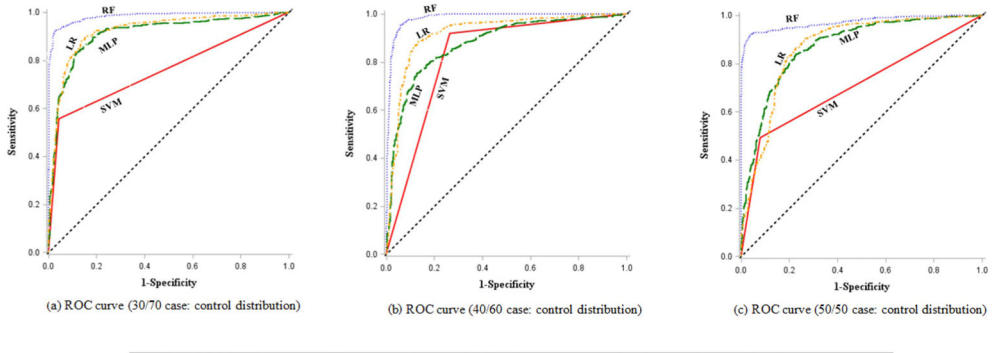


Fig. 1.
Shows the study flow.

81-90% (65 papers)	71-80% (51-54 papers)	61-70% (44-47 papers)	51-60% (35-40 papers)	41-50% (30-33 papers)
Age (65 papers)	BMI (54 papers) Family History of Diabetes (53 papers) Fasting/Blood/Plasma Gender (51 papers)	H/O Hypertension (47 papers) Blood pressure (44 papers)	Triglycerides (40 papers) Tobacco use (39 papers) H/O Diabetes (38 papers) Waist measurement (36 papers) H/O Hyperglycemia (35 papers)	Height (35 papers) Ethnicity (51 papers) Physical Activity (50 papers) Weight (30 papers) Dyslipidemia (30 papers) Hypertension modifications (30 papers)
31-40% (21-24 papers)	**21-30% (14-20 papers)**	**11-20% (7-13 papers)**	**5-10% (3-6 papers)**	
Race (24 papers) HbA1C (23 papers) Total cholesterol (22 papers) Diabetes medications (21 papers)	Oral glucose tolerance test (20 papers) Cardiovascular disease (20 papers) Insulin levels (19 papers) Educational level (17 papers) Alcohol consumption (16 papers) Low density lipids (14 papers)	Statins (13 papers) Myocardial infarction/ Stroke (11 papers) Household income (10 papers) Dietary (10 papers) Social class (4 papers) Gestational diabetes (8 papers) Metabolic Syndrome (8 papers) Fasting insulin (7 papers) Corticosteroid use (7 papers) Uric acid (7 papers)	C-reactive protein (6 papers) Have primary care provider (5 papers) Hepatic enzymes (5 papers) Waist to Hip ratio (4 papers) Insurance status (4 papers) Marital status (4 papers) Random Capillary Glucose (4 papers)	Insulin resistance (3 papers) 2h plasma insulin (3 papers) Creatinine levels (3 papers) Family History of Hypertension (3 papers) Baby weight (3 papers) White Blood Count (3papers)
2-4% (2 papers)		**0-2% (1 paper)**		
Heart rate Hip circumference Waist to height ratio Body fat distribution Percentage of body fat Insulin sensitivity	Polycystic Ovarian Syndrome Adiponectin Number of health visits Skin fold thickness C-peptide	Access to medical care ARIC diabetes risk score Breathlessness eGFR Erythrocyte count Ferritin levels FINDRISC score Forehead to neck distance Forehead to rib distance Forehead to waist distance Frequency of thirst, urination, tiredness and repeated cystitis	Geographic location H/O Hypercholesterolemia Hematocrit value 79. High albumin/creatinine ratio Hyperdynamic circulation Medical history Menopause Neck circumference Neck to hip distance Triacylglycerol levels	Obesity Observation Other diseases not specified otherwise Pregnancy Random blood loss Thrombocytes Times since last medical visit

Fig. 2.

Catalogs all retained variables identified across the 69 studies that met eligibility for data abstraction and were included for modeling.



Case: Control distribution	*MLP*	*LR*	*RF*	*SVM*
30-70	0.894	0.935	0.983	0.809
40-60	0.9	0.935	0.978	0.806
50-50	0.898	0.866	0.972	0.775

Area Under Curve (AUC) measures for all models.

Fig. 3.
Shows ROC (AUC) of all the four classifiers with varied case-control distribution.

Table 1

Illustrates list of all data features included in the prediction model development along with baseline characteristics.

Feature number	Feature	Cases (High Risk)	Controls (Low Risk)	P-Value
1	Age			
	21–30 years	442 (08.36%)	7488 (21.25%)	<0.0001
	31–40 years	769 (14.54%)	8236 (23.37%)	
	41–50 years	1245 (23.55%)	5940 (16.85%)	
	51–60 years	1434 (27.12%)	6587 (18.69%)	
	61–70 years	983 (18.59%)	4217 (11.96%)	
	71–80 years	350 (06.62%)	1923 (05.45%)	
	80–89 years	63 (01.19%)	842 (02.38%)	
2	**Body Mass Index (BMI) [25]** Less than 18.5 = Underweight; 18.5–24.99 = Normal; 25.00–29.99 = Overweight; 30.0 and above = Obese			
	Missing values	49 (00.93%)	2513 (07.13%)	<0.0001
	Underweight	25 (00.47%)	466 (01.32%)	
	Normal	284 (05.37%)	7946 (22.55%)	
	Overweight	933 (17.65%)	9402 (26.69%)	
	Obese	3995 (75.58%)	14,906 (42.31%)	
3–30	**Bleeding on probing (BOP)** ** Each tooth is probed at six sites.			
	Missing values (excludes extracted teeth)	117,488 (80.89%)	317,100 (35.40%)	<0.0001
	Total number of teeth with BOP present	8286 (05.71%)	172,834 (19.29%)	
	Total number of teeth with BOP absent	16,764 (11.54%)	318,382 (35.34%)	
31.	**Corticosteroids medications** (Retrieved from medication lists from the iEHR)			
	Missing values	0	477 (01.35%)	<0.0001
	Corticosteroid prescribed	1053 (19.92%)	7258 (20.32%)	
	Corticosteroid not prescribed	4233 (80.08%)	27,598 (78.33%)	
32.	**Serum Creatinine Levels** **Females:** Less than 0.6 mg/dl = low, 0.6 mg/dl to 1.1 mg/dl = Normal, More than 1.1 mg/dl = High **Males:** Less than 0.7 mg/dl = low,0.7 mg/dl to 1.3 mg/dl = Normal, More than 1.3 mg/dl = High			
	Missing values	1103 (20.87%)	25,124 (71.31%)	<0.0001
	Low	211 (03.99%)	292 (00.83%)	
	Normal	3670 (69.43%)	9271 (26.31%)	
	High	302 (05.71%)	546 (01.55%)	
33.	**Use of Diabetic Medications**			
	Missing	0	477 (01.35%)	<0.0001
	DM medication prescribed	1870 (35.38%)	1684 (04.78%)	
	DM medication not prescribed	3416 (64.62%)	33,072 (93.87%)	
34.	**Ethnicity**			

Feature number	Feature	Cases (High Risk)	Controls (Low Risk)	P-Value
	Missing	92 (01.74%)	3013 (8.55%)	<0.0001
	Declined	49 (00.93%)	352 (01.00%)	
	Hispanic or Latino	162 (03.06%)	113 (03.21%)	
	Not Hispanic or Latino	4975 (94.12%)	30,674 (87.06%)	
	Patient Does Not Know	8 (00.15%)	64 (00.18%)	
35.	**Family history of Diabetes** Family history included parents and siblings			
	Yes	459 (08.68%)	178 (00.51%)	<0.0001
	No	4827 (91.32%)	35,055 (99.49%)	
36.	**Gender**			
	Male	2443 (46.2%)	14,921 (42.3%)	<0.0001
	Female	2843 (53.7%)	20,312 (57.6%)	
37.	**High Density Lipids (HDL) cholesterol** Less than 40 mg/dl = Poor, 40 mg/dl to 59 mg/dl = Better, 60 mg/dl and above = Best			
	Missing	1516 (28.68%)	28,790 (81.71%)	<0.0001
	Poor	1749 (33.09%)	1347 (03.82%)	
	Better	1711 (32.37%)	3192 (09.06%)	
	Best	310 (05.86%)	1904 (05.40%)	
38.	**Hypertension** <120 mm Hg (SBP) and <80 mm Hg (DBP) = Normal; 120–129 mm Hg (SBP) and <80 mm Hg (DBP) = Prehypertension; 130–139 mm Hg (SBP) or 80–89 mm Hg (DBP) = Stage 1 hypertension; ≥140 mm Hg (SBP) or ≥90 mm Hg (DBP) = Stage 2 hypertension; ≥180 mm Hg (SBP) or ≥120 mm Hg (DBP) = Hypertensive crisis			
	Missing	1453 (27.49%)	1551 (04.40%)	<0.0001
	Normal	1084 (20.51%)	11,896 (33.76%)	
	Prehypertension	653 (12.35%)	6287 (17.84%)	
	Stage 1 hypertension	1250 (23.65%)	11,439 (32.47%)	
	Stage 2 hypertension	824 (15.59%)	3976 (11.28%)	
	Hypertensive crisis	22 (00.42%)	84 (00.24%)	
39.	**Use of Hypertensive medications**			
	Missing	0	477 (01.35%)	<0.0001
	Hypertensive medication prescribed	2001 (37.85%)	9026 (25.62%)	
	Hypertensive medication not prescribed	3285 (62.15%)	25,730 (73.03%)	
40–45.	**Insurance** Types: Medicare, Medicaid, Commercial, Self-pay, Senior Care (Prescription only), No health insurance			
	Yes	5195 (98.28%)	33,483 (95.03%)	<0.0001
	No	91 (01.72%)	1750 (04.97%)	
46.	**LDL cholesterol** Less than 100 mg/dl = Optimal, 100 mg/dl to 129 mg/dl = Near optimal, 130 mg/dl to 159 mg/dl = Borderline high, 160 mg/dl to 189 mg/dl = High, 190 mg/dl and above = Very High			
	Missing	1666 (31.52%)	28,855 (81.90%)	<0.0001
	Optimal	1786 (33.79%)	2766 (07.85%)	
	Near optimal	1082 (20.47%)	2111 (05.99%)	
	Borderline high	528 (09.99%)	1078 (03.06%)	
	High	165 (03.12%)	311 (00.88%)	

Feature number	Feature	Cases (High Risk)	Controls (Low Risk)	P-Value
	Very High	59 (01.12%)	112 (00.32%)	
47.	**Total number of unique dental visits in the given measurement year**			
	Continuous variable	1.08 ± 2.64	3.49 ± 2.66	<0.0001
48.	**Periodontal Disease (PD) Types**			
	Missing	2300 (43.51%)	8783 (24.93%)	<0.0001
	Healthy	78 (01.48%)	595 (01.69%)	
	Type 1	323 (06.11%)	5370 (15.24%)	
	Type 2	1882 (35.60%)	16,297 (46.25%)	
	Type 3	623 (11.79%)	3712 (10.54%)	
	Type 4	78 (01.48%)	445 (01.26%)	
	Type 5	2 (00.04%)	31 (00.09%)	
49–76.	**Periodontal Pocket Depth (PPD)** Each tooth is probed at six sites and maximum PPD value is assigned as the PPD for each tooth.			
	Total number of teeth with missing PPD values (excludes extracted teeth)	120,517 (82.98%)	354,068 (39.53%)	<0.0001
	Total number of teeth with PPD > 5 mm	1270 (05.14%)	17,794 (03.28%)	
77–84.	**Race** **Non-White:** American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander; **White**; **Unknown**: Patient Does Not Know, Declined and Unknown			
	White	4952 (92.87%)	30,363 (85.57%)	<0.0001
	Non-White	206 (03.86%)	1328 (03.74%)	
	Unknown	174 (03.26%)	3791 (10.68%)	
85.	**Use of Statins**			
	Missing	0	477 (01.35%)	<0.0001
	Statin prescribed	1342 (25.39%)	4941 (14.02%)	
	Statin not prescribed	3944 (74.61%)	29,815 (84.62%)	
86.	**Tobacco use status**			
	Missing	2452 (46.39%)	2393 (06.79%)	<0.0001
	Current user	847 (16.02%)	13,041 (37.01%)	
	Former user	809 (15.30%)	7712 (21.89%)	
	Never	1178 (22.29%)	12,087 (34.31%)	
87–114.	**Total number of missing teeth**			
	Total number of missing teeth (includes extracted teeth)	2767 (01.87%)	90,756 (09.2%)	<0.0001
115.	**Total Triglycerides** Less than 150 mg/dl = Normal, 150 mg/dl to 199 mg/dl = Borderline high, 200 mg/dl to 499 mg/dl = High, 500 mg/dl and above = Very High.			
	Missing	1532 (28.98%)	28,805 (81.76%)	<0.0001
	Less than 150 mg/dl (Normal)	1681 (31.80%)	4628 (13.14%)	
	150 mg/dl to 199 mg/dl (Borderline high)	810 (15.32%)	878 (02.49%)	
	200 mg/dl to 499 mg/dl (High)	1150 (21.76%)	875 (02.48%)	
	500 mg/dl and above (Very High)	113 (02.14%)	47 (00.13%)	

Feature number	Feature	Cases (High Risk)	Controls (Low Risk)	P-Value
116.	**WBC** Less than $4.0*10^9$/L = Leukopenia, $4.0*10^9$/L to 11.0 $*10^9$/L = Normal, More than $11.0*10^9$/L = Leukocytosis			
	Missing	1883 (34.68%)	25,958 (73.68%)	<0.0001
	Less than 4.0*109/L (Leukopenia)	33 (00.62% 0	233 (00.66%)	
	4.0*109/L to 11.0 *109/L(Normal)	2944 (55.69%)	7869 (22.33%)	
	More than 11.0*109/L (Leukocytosis)	476 (09.00%)	1173 (3.33%)	

Table 2

List of all data features deleted and corresponding percentage of missing values.

Feature	Description	Cases (% of missing values(N = 5319)	Controls (% of missing values(N = 36,224)
CRP	C-reactive protein (Continuous value)	99.4	99.9
Bone loss	Presence/absence of periodontal bone loss (Boolean value)	98.8	96.9
Plaque	Presence/absence of plaque (Boolean value)	98.7	96.7
Mobility	Presence/absence of tooth mobility (Boolean value)	97.3	94.4
Uric Acid	Uric acid levels(Continuous value)	90.6	98.9
Gingivitis	Presence/absence of gingivitis(Boolean value)	83.1	0
Xerostomia	Presence/absence of xerostomia (Boolean value)	83.1	0
Oral candidiasis	Presence/absence of oral candidiasis (Boolean value)	83.1	0

Table 3

Shows Case-control distribution of training/testing datasets.

Case-Control	Cases[a]	Controls[b]
30–70	4757	11,100
40–60	4757	7136
50–50	4757	4757

[a]Number of observations kept consistent.

[b]Number of observations sampled without replacement.

Table 4

Confusion matrix of the best performing model (RF with 50-50(case: control)).

Total Population ($N_{validation}$) N = 1058	**Actual Positive**	**Actual Negative**	**ACC = 0.9414 (ΣTP+ΣTN /ΣTotal Population)**
Predicted Positive	TP = 481	FP = 48	PPV = 0.943 (ΣTP/ΣPredicted Positive)
Predicted Negative	FN = 14	TN = 515	NPV = 0.974 (TN/ΣPredicted Negative
	TPR = 0.972 (ΣTP/ΣActual Positive)	FPR = 0.085 (ΣFP/ΣActual Negative)	
	FNR = 0.028 (ΣFN/ΣActual Positive)	TNR = 0.915 (ΣTN/ΣActual Negative)	

TP = True Positive; TN = True Negative; FP=False Positive; FN=False Negative; ACC = Accuracy; TPR = True positive rate; TNR = True negative rate; FPR=False positive rate; FNR=False negative rate; PPV=Positive predictive value; NPV=Negative predictive value.

Table 5

Performance metrics for all models.

Case-Control	Classifier	Total accuracy	AUC	MCC	F-measure	Sensitivity	Specificity	Precision
30–70	MLP	81.29%	0.894	0.628	0.813	0.813	0.789	0.815
	LR	81.19%	0.935	0.654	0.808	0.812	0.74	0.843
	RF	91.68%	0.983	0.842	0.916	0.917	0.864	0.926
	SVM	80.91%	0.809	0.65	0.804	0.809	0.736	0.842
40–60	MLP	82.51%	0.900	0.65	0.825	0.825	0.826	0.825
	LR	86.29%	0.935	0.734	0.862	0.863	0.816	0.871
	RF	92.72%	0.978	0.860	0.927	0.927	0.883	0.933
	SVM	85.35%	0.806	0.719	0.852	0.853	0.799	0.865
50–50	MLP	82.42%	0.898	0.652	0.824	0.824	0.792	0.828
	LR	75.24%	0.866	0.519	0.749	0.752	0.704	0.767
	RF[a]	94.14%	0.972	0.885	0.941	0.972	0.915	0.943
	SVM	77.50%	0.775	0.581	0.769	0.775	0.708	0.807

[a]Best performing model.