HHS Public Access

Author manuscript

Gerodontology. Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

Gerodontology. 2019 December ; 36(4): 395–404. doi:10.1111/ger.12432.

Application of machine learning for diagnostic prediction of root caries

Man Hung, PhD[1,2,3,4,5], **Maren W. Voss, ScD**[2], **Megan N. Rosales, MS**[2], **Wei Li, MS**[4], **Weicong Su, MS**[2], **Julie Xu, BA**[2], **Jerry Bounsanga, BS**[2], **Bianca Ruiz-Negrón**[2], **Evelyn Lauren**[2], **Frank W. Licari, DDS**[1]

[1]Roseman University of Health Sciences College of Dental Medicine

[2]University of Utah Department of Orthopaedic Surgery Operations

[3]University of Utah Study Design and Biostatistics Center

[4]University of Utah Department of Family and Preventive Medicine

[5]Huntsman Cancer Institute

Abstract

Objective: This study sought to utilize machine learning methods in artificial intelligence to select the most relevant variables in classifying the presence and absence of root caries and to evaluate the model performance.

Background: Dental caries is one of the most prevalent oral health problems. Artificial intelligence can be used to develop models for identification of root caries risk and to gain valuable insights, but it has not been applied in dentistry. Accurately identifying root caries may guide treatment decisions, leading to better oral health outcomes.

Methods: Data were obtained from the 2015–2016 National Health and Nutrition Examination Survey and were randomly divided into training and test sets. Several supervised machine learning methods were applied to construct a tool that was capable of classifying variables into the presence and absence of root caries. Accuracy, sensitivity, specificity and area under the receiver operating curve were computed.

Correspondence: Man Hung, PhD, Professor and Research Dean, College of Dental Medicine, Roseman University of Health Sciences, 10894 South River Front Parkway, South Jordan, UT 84095, 801-878-1270, mhung@roseman.edu.

Author Contributions

Man Hung: conception and design of the study, data acquisition, analysis and interpretation of data, drafting and revising the article for important intellectual content, and final approval of the manuscript.
Maren W. Voss: design of the study, data acquisition, interpretation of data, drafting the article and final approval of the manuscript.
Megan N. Rosales: design of the study, data acquisition, analysis and interpretation of data, revising the article, and final approval of the manuscript.
Wei Li: design of the study, acquisition of data, analysis of data, revising the article, and final approval of the manuscript.
Weicong Su: design of the study, analysis and interpretation of data, revising the article, and final approval of the manuscript.
Julie Xu: design of the study, interpretation of data, drafting the article and final approval of the manuscript.
Jerry Bounsanga: design of the study, acquisition of data, revising the article, and final approval of the manuscript.
Bianca Ruiz-Negrón: design of the study, interpretation of data, revising the article, and final approval of the manuscript.
Evelyn Lauren: design of the study, interpretation of data, revising the article, and final approval of the manuscript.
Frank W. Licari: conception and design of the study, interpretation of data, revising the article for important intellectual content, and final approval of the manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest.

Results: Of the machine learning algorithms developed, support vector machine demonstrated the best performance with an accuracy of 97.1%, precision of 95.1%, sensitivity of 99.6%, and specificity of 94.3% for identifying root caries. The area under the curve was 0.997. Age was the feature most strongly associated with root caries.

Conclusion: The machine learning algorithms developed in this study perform well and allow for clinical implementation and utilization by dental and non-dental professionals. Clinicians are encouraged to adopt the algorithms from this study for early intervention and treatment of root caries for the aging population of the United States, and for attaining precision dental medicine.

Keywords

Dental medicine; quality of life; machine learning; artificial intelligence; root caries; National Health and Nutrition Examination Survey (NHANES)

Introduction

Dental caries is one of the most prevalent oral health problems[1] and is considered to be one of the most important and preventable global oral health concerns.[2] Dental caries is defined as the localized destruction of susceptible dental hard tissues by acidic by-products from bacterial fermentation of dietary carbohydrates and can occur on the crown (coronal caries) and at the root (root caries) of the tooth.[3] According to the National Institute of Dental and Craniofacial Research, 92% of adults aged 20 to 64 years in the United States had some experience with dental caries in their permanent teeth in 2011–2012, while the percent rose to 96% for adults aged 65 and over.[4,5] If left untreated, caries can lead to tooth loss[6,7] as well as reduced quality of life,[8] ease of daily living, and self-concept regarding their oral health.[9] Thus, minimizing the experience and impact of caries on individuals' general health and quality of life is an important public health issue.

Research into the prevalence and impact of root caries has demonstrated that a number of individual factors are related to poor oral health. Males show a higher prevalence of untreated caries than females.[10] Individuals belonging to racial/ethnic minority groups, such as Native Americans, Blacks, and Hispanics, have a higher prevalence of periodontal diseases, untreated root caries, tooth loss, and generally experience a higher incidence of oral cancer than non-Hispanic Whites.[11] Socioeconomic status components such as income,[10,12] living condition,[13] education,[10,14] and access to dental care[10,14] are all factors that can contribute to dental disparities. With the inability to access to dental care, 21% of Latino children under age 17 are without dental insurance compared to just 6% uninsured rates for whites and 7% for African Americans.[15] The issue of accessibility becomes more prominent as Medicare, for those 65 and older, doesn't provide for routine dental care, and older adults may experience higher rates of tooth decay than children.[16] Lifestyle factors such as poor diet, nutrition, and a lack of dental hygiene play key roles in disparities as well.[2] Among the vulnerable, elder population, root caries tend to occur due to reduced upkeep of dental hygiene practices.[17] While a large portion of the currently affected population can retain their teeth for a majority of their lifespan with simple individual and population-level interventions, such as water fluoridation and regular professional preventive dental care,[18]

socioeconomic components remain a large contributor to increased prevalence of poor oral health.

Looking specifically at dental factors that are associated with root caries, self-reported dry mouth,[19] number of teeth at baseline,[20] and gingival recession[21,22] have been associated with root caries. Among elders, surfaces with visible plaque, denture contact, and more prominent gingival recession are areas that are likelier to get affected by root caries[23]. However, generally with those older than 35 years, complexities arise in determining relationships with root caries as the presence of periodontal disease increases and becomes the primary culprit of tooth loss.[18]

While incredibly important, prevalence and outcome research can often be difficult to use when attempting to develop clinical interventions.[22,24] Consequently, we believe there can be clinical benefits from employing artificial intelligence to the prediction of root caries. Machine learning methods in artificial intelligence have been previously applied to different areas of healthcare and have the ability to explore large amounts of data to reveal patterns and complex relationships between variables.[25,26] They have strong potential to produce precise and individualized prediction of root caries risk.

To our knowledge, machine learning has not been used to develop models in identification of root caries risk. This study utilized machine learning techniques to identify the likelihood of a person to develop root caries by selecting the most relevant variables from demographic and lifestyle factors. A potential application of this technique is to detect root caries on an individual level, enabling evidence-based personalized dental medicine that may assist in decreasing root caries experience of aging populations via early prevention and treatment.

Materials and Methods

Data.

This study used public data from the National Health and Nutrition Examination Survey (NHANES) 2015–2016 cycle.[27] Since these were de-identified, public data available from the NHANES website, Institutional Review Board (IRB) approval was not required. This study was considered exempt from IRB evaluation on the basis of federal regulation 45 CFR 46.101(b) (research involving the study of secondary data recorded in such a manner that subjects cannot be identified). The NHANES is a study of the National Center for Health Statistics within the Centers for Disease Control and Prevention. It is conducted annually via both interviews and clinical examinations to assess the health status of adults and children in the United States. It includes information derived from questionnaires on demographics, socioeconomic status, dietary, and health-related topics. Additionally, the NHANES has a clinical examination component which includes medical, dental, and physiological measures.

For the 2015–2016 cycle, the NHANES included oversampling of underrepresented groups. A total of 15,327 people was invited to participate in the study and of those invited, 9,971 people completed the interview and 9,544 people were examined. Interview questions were administered by trained interviewers in the participant’s home, but sensitive questions

regarding alcohol use and reproductive health were administered at the examination center. Clinical examinations were conducted at a mobile examination center at designated locations by licensed and trained medical personnel.

Outcome Variable.

The oral health outcome variable of interest for this study was root caries. It was a dichotomous variable with either yes or no to indicate the presence or absence of one or more root caries based on clinical examination. Dental caries was defined as the localized destruction of susceptible dental hard tissues by acidic by-products from bacterial fermentation of dietary carbohydrates occurring either on the crown or root of the tooth.[3,28] This study focused on root caries as it is a more serious condition that can lead to greater oral health issues but is also highly treatable and can be prevented.[29] Root caries was identified in oral examinations by licensed and trained dental professionals from NHANES during a dental caries assessment using a decayed, missing, and filled surface index. The presence of root caries was defined as the presence of one or more untreated (decayed, D-root) root caries lesions or treated (filled, F-root) root surfaces. The outcome variable used in this study included the presence/absence of D-root and/or F-root.

Analytical Approach.

Demographic and clinical characteristics of the participants were examined in terms of mean, standard deviation, frequency and proportion where appropriate. Machine learning methods were utilized to classify the presence or absence of root caries. In machine learning, computer algorithms can be applied to a training data set. These algorithms “learn” the patterns which are present in the data and automatically generate rules that are used to conduct data mining or predict future outcome from the features (i.e., variables). These predictions can then be compared against the actual values from a test data set (i.e., validation data set) to assess the performance of the machine generated rules. Machine learning is particularly helpful when dealing with large and complex data where the relationships between variables are not obvious. It is useful for clinical decision support and can contribute to diagnosis and prognosis of oral health conditions as well as personalized or individualized dental treatment regimes.

There was a total of 9,971 cases and 950 variables present in the complete dataset. To prepare the data for processing, all cases with missing data for root caries as well as variables that had 50% or more missing data were excluded, resulting in 357 variables and a total sample size of 5,135. To minimize bias and enhance efficiency of interpretation, variables that were unlikely to be related to root caries (e.g., subject IDs) and variables providing essentially the same information (e.g., age in a continuous scale and age in categorical groupings) as well as variables that were the likely results of possessing root caries (e.g., recommendation for dental care) were removed. The resulting variables were subjected to independent samples t-tests for continuous variables and chi-square tests for categorical variables to examine whether there were significant differences between the root caries and no root caries groups. A total of 37 variables demonstrated statistically significant relationships with the outcome variable root caries ($p<0.001$) (Appendix A). These 37 variables were inputted into initial machine learning models to determine their relative

importance based on their F-scores. The F-score is a measure that determines feature importance based on how often that feature is taken into account during the machine learning process. Variables with higher F-scores contributed more to the prediction of root caries. In order to achieve parsimony, the top 15 most important variables were selected to construct machine learning models. The data were then randomly partitioned into training and test sets with 80% for training and 20% for testing. Since the original data were highly imbalanced (containing 4,344 cases without root caries but only 791 cases with root caries), sampling with replacement, specifically oversampling, was used to create balanced data for the under-represented class (i.e., minority class). The balanced data contained 4,746 cases with root caries and 4,344 cases without root caries. Altogether, a total of 9,090 cases were used for training and testing in machine learning, with 7,272 cases (80% of 9,090) randomly selected for training and 1,818 cases (20% of 9,090) for testing.

Imbalanced data are known to introduce a high degree of classification bias to model performance (e.g., sensitivity, specificity) such that the machine learning algorithms are almost never be able to predict the minority class, and that the majority class almost always has inflated model performance. Thus, when using highly imbalanced data set, we often see a large gap between sensitivity and specificity of machine learning models and a high misclassification rate for the minority class.[30,31] In order to solve such issues, various strategies such as oversampling or undersampling have been proposed to reduce the inherent bias resulting from imbalanced data. Oversampling has demonstrated to be able to reduce the gap between sensitivity and specificity and lower the misclassification rate for the minority class.[30,31] On the other hand, if not done properly, oversampling can result in overfitting issues such as obtaining perfect accuracy and AUC when in reality they are not perfect. In this study, we strived to minimize overfitting issues by using a separate validation sample for model validation.

Several supervised machine learning methods were applied to generate the prediction of root caries for individuals. These include support vector machine (SVM), extreme gradient boosting (XGBoost), random forest regression (RF), k-nearest neighbors (k-NN), and logistic regression.[32–35] Logistic regression was chosen because it was commonly used in traditional medical studies; all other methods were chosen due to their tolerance to overfitting, ability to model nonlinear relationships, ease for implementation in clinical settings, or acceptability in the machine learning community.

These machine learning algorithms were coded using Python 3.7.0 (Python Software Foundation) and WEKA 3.8.2 (University of Waikato, Hamilton, New Zealand). The test dataset was used to compute accuracy, sensitivity, specificity, and area under the curve (AUC) of the receiver-operating characteristic (ROC) curve. Accuracy of the prediction was considered as the most relevant for clinical applications in dental care.

Results

The sample size for this study was 5,135. Males made up 48.4% of the sample, and females made up 51.6%. A total of 1,629 (31.7%) identified as White or Caucasian, 1,094 (21.3%) as black or African American, 1,613 (31.4%) as Hispanic or Mexican American, and 611

(11.9%) as Asian. The average age of the sample was 46.6 (standard deviation = 18.1; median = 46.0) (Table 1).

Figure 1 displays a visual presentation of variable importance, reflecting the contribution of each of the thirty-seven significant indicators of root caries to the machine learning model. The larger the F-score of a variable, the higher the contribution it has on the identification of root caries. Age was found to be the most important variable in identifying root caries. The top fifteen features included five demographic variables (i.e., age, household income, education, race/ethnicity, and marital status), five oral health variables (i.e., last dentist visit, flossing, mouth ache, self-rated oral health, and oral embarrassment), and five lifestyle/ health variables (i.e., TV watching, computer use, use of sunscreen, alcohol consumption, and cholesterol prescriptions) (Figure 1 and Table 2). The F-scores calculated by the various machine learning algorithms were slightly different, but this difference was minor, and the rankings of the variables remained relatively consistent.

Classification results for the machine learning algorithms are presented in Table 3. The top classifier was SVM with an AUC of 0.997, an accuracy of 97.1%, precision of 95.1%, sensitivity of 99.6% and specificity of 94.3% for identification of root caries. The XGBoost and RF also performed very well with an overall accuracy of 94.7% and 94.1% respectively. The kNN was satisfactory at 83.2% accuracy. The commonly used logistic regression in traditional research studies performed the worst relative to the other algorithms in this study but still had a reasonable accuracy of 74.3%.

Figure 2 displays a graphical plot of the ROC curves for all of the machine learning algorithms utilized in this study. The AUC for the logistic regression was adequate at 0.818. The SVM, XGBoost and RF had an AUC of 0.997, 0.987, and 0.999, revealing exceptionally high model performance.

Discussion

Root caries is a significant public health concern and has been increasing in prevalence. The use of machine learning to identify factors related to root caries is an opportunity to improve oral health with consequent effects on general health. This is the first study using machine learning methods in artificial intelligence to identify root caries from a large scale of data consisting of demographic, nutrition, lifestyle, laboratory, and oral examination variables. We used the NHANES data and applied multiple machine learning methods to identify the best model and factors related to root caries. The best performing machine learning model was SVM, which most accurately classified the presence versus absence of root caries.

Across all methods, four variables were consistently identified as the most critical in indicating the presence of root caries. These were age, income, date of last dental visit, and hours of television watching. Age was the most relevant predictive variable, consistent with evidence that root caries increases with age due to increasing exposure of root surfaces among other things.[36] Low income as a factor of socioeconomic status[10,12] and as an indicator of a financial barrier[10,14] to dental care access has also been associated with oral health disparities. Receiving dental care from a professional on a regular basis increases

chances of early diagnosis, prevention and treatment of oral diseases.[37,38] Consequently, previous research has shown that those who do not receive regular care have worse oral health than those who do.[39] Last dental visit as a prominent oral health feature is especially consistent with the idea of reduced accessibility and increased prevalence among elders. Overall, this confirms previously identified features. Yet, the value of the machine learning approach comes with the identification of unexpected and less intuitive features, such as hours spent watching television, as important indicators of root caries risk. While lifestyle factors in general may not be directly responsible for the development of root caries, they may provide an indirect link to a person's overall health, lifestyle, and likelihood of developing oral health problems.

Ultimately, most of our other top features were consistent with prior research that has identified demographic, lifestyle, and oral health variables as important features of poor oral health. We found that education, marital status, race/ethnicity and gender and demographic factors were indicative of root caries. Meanwhile previous research has identified gender,[10] ethnicity,[11] socioeconomic status,[10,12] living conditions,[13] and education[10,14] as factors contributing to oral health disparities. We found high alcohol consumption as an indicator of root caries, and although an association has yet to be made with root caries specifically, high alcohol consumption has been found to be associated with larger amount of caries on tooth surfaces.[40] Finally, we identified four other oral health variables relevant in classifying root caries: aching in mouth, self-rated oral health, flossing, and oral embarrassment. Although literature on the effects of flossing on dental health is inconclusive,[41,42] previous research has cited correlation of poor oral health with aching in mouth,[43] oral embarrassment,[44] and self-rated oral health.[45]

Sunscreen use, computer use, and taking prescription medicine for cholesterol were some of the unique indicators we discovered, which did not exist in the current literature. Similar to the case with hours spent watching television, factors like use of sunscreen and computer use, although not directly indicative, may provide insight to a patient's oral health practices and habits. Taking prescription medicine for cholesterol may suggest why prior evidence shows an association between dental health and heart disease/elevated cholesterol.[46,47] Since commonly prescribed cholesterol lowering drugs such as anticholinergics decrease salivary gland function,[48] perhaps patients with heart disease/elevated cholesterol are at an increased risk for root caries due to their medications.

This study was not without limitations. First, we utilized a large amount of data collected by NHANES from a large sample in the United States. The findings derived from this large sample is meant to be more representative of and can be generalizable to the United States' population. Yet individual dental clinics may have different patient demographics and may exhibit different characteristics. Second, in machine learning a large amount of data or variables is often used in search for novel insights, which makes statistical significance testing inapplicable or losing its meaning. However, since a central aim of this study is discovery and exploration of actionable new insights, not statistical significance testing, applying a large number of variables is not of concern but of great benefit for building accurate models in artificial intelligence. Third, the machine learning feature selection did not account for the covariance between lifestyle factors and the oral hygiene variables. By

only using F-score to select feature importance, variables that may not have been directly correlated with root caries, but rather were associated with other variables that influence root caries may have been selected. This may have been the case with features such as age and taking cholesterol medications, and some of the lifestyle factors. In the future, it may be beneficial to compare the current models against other machine learning models in predicting root caries that use different methods of feature selection. Fourth, this was a cross-sectional study, so the final model showed possible indicators associated with the presence or absence of root caries. Longitudinal study is needed in the future to establish and confirm the predictive ability of the model. Additionally, onsite clinical validation has not been started but future research can focus on such validation to improve the algorithms.

While confirming prior research regarding significant indicators of root caries at a population level, our study also developed highly accurate and precise computer algorithms to model risk for individual patients. The application of machine learning in artificial intelligence not only approximated dentists' examination skills, but discovered novel and complex relationships not readily apparent to dentists or humans in general. The use of machine learning methods did not simply help us in identification of risk factors for root caries, it helped us to generate computer algorithms that are able to consider combinations of variables to classify the presence and absence root caries. Discovering and incorporating such combinations of variables and their complex relationships with root caries to guide understanding and individual treatment decision can be a challenge for humans, but they can be a reality with artificial intelligence (such as Alexa, self-driven cars, face ID for unlocking phones, or other robots that we have seen and used in our daily life). Machine learning is the driver of artificial intelligence and has powerful public health implications when applied to clinical problems.

Innovations using artificial intelligence have the ability to disrupt and advance the areas of diagnosis and prognosis in oral health. In the future, real-time online clinical decision support tool can be made by incorporating the machine learning algorithms developed from this study to facilitate precision medicine in oral care. This can be used as a screening tool in general medical practices, dental clinics, social service centers, or placed online, providing recommendations for dental examinations for those identified at high risk. The information derived from the machine learning findings in this study also included the identification of other medical conditions or life styles to the presence of root caries, which is probably more applicable to be utilized by non-dental professionals to categorize patients that might be of higher risk to develop root caries and provide referrals of those patients to oral health professionals for further evaluation and early intervention and prevention.

According to the U.S. Census Bureau's 2017 National Population Projections, the size of the older population will expand by 2030 such that 1 in every 5 people will be at the retirement age of 65 or older.[49] With an increasingly aging population, root caries and other oral health outcomes that most commonly affect the elder population will only increase in prevalence. Therefore, the use of machine learning methods to understand root caries represents an incredible opportunity for early intervention and the improvement of oral health for the aging population. This is the first study applying machine learning to classify root caries and it has generated highly robust and accurate computer algorithms. The use of these algorithms

may enable the development of automated and cost-efficient tools for dental care and precision medicine and may have huge implications in intervention for those that are or could be affected by root caries and other oral health conditions.

Conclusion

Root caries is a considerably prevalent oral health problem; therefore, developing models that can inform diagnostic decisions or preventive measures on root caries has significant health benefits. In this study, we explored features that indicate root caries occurrence. The work presented here demonstrated a clear potential for the application of machine learning methods to identify hidden features that had never been known. The models developed in this study showed high accuracy, sensitivity, specificity, precision and AUC in classifying the presence and absence of root caries.

Acknowledgements

This project was supported by the Roseman University College of Dental Medicine Clinical Outcomes Research and Education (http://codmresearch.com), the University of Utah Undergraduate Research Opportunity Program, and the Population Health Research Foundation with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5UL1TR001067–02.

Appendix A:: Indicators of root caries

Variable Name	Variable Description	N	Root Caries/Yes	Root Caries/No	P value
BPQ090D	Told to take prescription for cholesterol	3,828	592	3,236	<.001
DEQ034D	Use sunscreen?	3,397	439	2,958	<.001
DMDEDUC2	Education Level - Adults 20+	4,873	789	4,084	<.001
DMDMARTL	Marital Status	4,874	790	4,084	<.001
INDHHIN2	Annual Household Income	4,801	732	4,069	<.001
OHQ030	When did you last visit a dentist	5,122	789	4,333	<.001
OHQ620	How often last year had aching in mouth?	3,962	731	3,231	<.001
OHQ680	Last year embarrassed because of mouth	3,966	731	3,235	<.001
OHQ845	Rate the health of your teeth and gums	5,133	791	4,342	<.001
OHQ870	How many days use dental floss/device	3,962	730	3,232	<.001
PAQ710	Hours watch TV or videos past 30 days	5,124	788	4,336	<.001
PAQ715	Hours use computer past 30 days	5,133	789	4,344	<.001
RIDAGEYR	Age in years at screening	5,135	791	4,344	<.001
RIDRETH1	Race/Ethnicity - Recode	5,135	791	4,344	<.001
SMQ020	Smoked at least 100 cigarettes in life	5,128	789	4,339	<.001
ALQ151	Ever have 4/5 or more drinks every day?	3,865	620	3,245	<.001
BPQ020	Ever told you had high blood pressure	5,133	790	4,343	<.001
DIQ010	Doctor told you have diabetes	5,132	790	4,342	<.001
DLQ020	Have serious difficulty seeing?	5,133	790	4,343	<.001
DLQ040	Have serious difficulty concentrating?	5,131	790	4,341	<.001

Variable Name	Variable Description	N	Root Caries/Yes	Root Caries/No	P value
DLQ050	Have serious difficulty walking?	5,135	791	4,344	<.001
DLQ060	Have difficulty dressing or bathing?	5,134	791	4,343	<.001
DPQ080	Moving or speaking slowly or too fast	4,717	733	3,984	<.001
IMQ011	Received hepatitis A vaccine	4,179	678	3,501	<.001
IMQ020	Received hepatitis B 3 dose series	4,331	691	3,640	<.001
INQ300	Family has savings more than $20,000	4,752	733	4,019	<.001
MCQ160B	Ever told had congestive heart failure	4,863	784	4,079	<.001
MCQ160F	Ever told you had a stroke	4,870	790	4,080	<.001
MCQ160O	Ever told you had COPD?	4,868	788	4,080	<.001
OHQ640	Last yr had diff w/ job because of mouth	3,965	731	3,234	<.001
OHQ770	Past yr need dental but couldn't get it	5,003	765	4,328	<.001
PAQ650	Vigorous recreational activities	5,134	790	4,344	<.001
PFQ049	Limitations keeping you from working	4,872	790	4,082	<.001
PFQ054	Need special equipment to walk	4,873	789	4,084	<.001
PFQ057	Experience confusion/memory problems	4,870	789	4,081	<.001
PFQ090	Require special healthcare equipment	4,874	790	4,084	<.001
RIAGENDR	Gender	5,135	791	4,344	<.001

References

1. Listl S, Galloway J, Mossey PA, Marcenes W. Global Economic Impact of Dental Diseases. Journal of dental research 2015;94(10):1355–1361. [PubMed: 26318590]
2. Petersen PE, Bourgeois D, Ogawa H, Estupinan-Day S, Ndiaye C. The global burden of oral diseases and risks to oral health. Bulletin of the World Health Organization 2005;83(9):661–669. [PubMed: 16211157]
3. Selwitz RH, Ismail AI, Pitts NB. Dental caries. The Lancet 2007;369(9555):51–59.
4. NIDOR. Dental Caries (Tooth Decay) in Adults (Age 20 to 64) National Insitute of Dental and Orofacial Research 2018(Accessed May 29, 2018):https://www.nidcr.nih.gov/research/data-statistics/dental-caries/adults#table1.
5. Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ. Prevalence of periodontitis in adults in the United States: 2009 and 2010. Journal of dental research 2012;91(10):914–920. [PubMed: 22935673]
6. Al-Shammari KF, Al-Khabbaz AK, Al-Ansari JM, Neiva R, Wang HL. Risk indicators for tooth loss due to periodontal disease. Journal of periodontology 2005;76(11):1910–1918. [PubMed: 16274310]
7. Akhter R, Hassan NM, Aida J, Zaman KU, Morita M. Risk indicators for tooth loss due to caries and periodontal disease in recipients of free dental treatment in an adult population in Bangladesh. Oral health & preventive dentistry 2008;6(3):199–207. [PubMed: 19119574]
8. Gerritsen AE, Allen PF, Witter DJ, Bronkhorst EM, Creugers NH. Tooth loss and oral health-related quality of life: a systematic review and meta-analysis. Health and quality of life outcomes 2010;8:126. [PubMed: 21050499]
9. Bennadi D, Reddy CVK. Oral health related quality of life. Journal of International Society of Preventive & Community Dentistry 2013;3(1):1–6. [PubMed: 24478972]
10. Gupta N, Vujicic M, Yarbrough C, Harrison B. Disparities in untreated caries among children and adults in the U.S., 2011–2014. BMC oral health 2018;18(1):30. [PubMed: 29510696]

11. Weatherspoon DJ, Chattopadhyay A, Boroumand S, Garcia I. Oral cavity and oropharyngeal cancer incidence trends and disparities in the United States: 2000–2010. Cancer Epidemiology 2015;39(4):497–504. [PubMed: 25976107]
12. Blas E, Kurup A. Priority public health conditions knowledge network of the Commission on Social Determinants of Health. In: Organization WH, ed. Equity, social determinants and public health programmes Geneva2010.
13. Eisen CH, Bowie JV, Gaskin DJ, LaVeist TA, Thorpe RJ Jr. The contribution of social and environmental factors to race differences in dental services use. Journal of urban health : bulletin of the New York Academy of Medicine 2015;92(3):415–421. [PubMed: 25680951]
14. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public health reports (Washington, DC : 1974) 2014;129 Suppl 2:19–31.
15. Flores G, Tomany-Korman SC. Racial and ethnic disparities in medical and dental health, access to care, and use of services in US children. Pediatrics 2008;121(2):e286–298. [PubMed: 18195000]
16. National Center for Chronic Disease Prevention and Health Promotion DoOH. Oral Health for Older Americans Oral Health 2013; https://www.cdc.gov/oralhealth/publications/factsheets/adult_oral_health/adult_older.htm, 2018.
17. Bignozzi I, Crea A, Capri D, Littarru C, Lajolo C, Tatakis DN. Root caries: a periodontal perspective. J Periodontal Res 2014;49(2):143–163. [PubMed: 23647556]
18. Schoen a MH, Freed JR. Prevention of Dental Disease: Caries and Periodontal Disease. Annual Review of Public Health 1981;2(1):71–92.
19. Chi DL, Berg JH, Kim AS, Scott J. Correlates of root caries experience in middle-aged and older adults in the Northwest Practice-based REsearch Collaborative in Evidence-based DENTistry research network. Journal of the American Dental Association (1939) 2013;144(5):507–516. [PubMed: 23633699]
20. Fure S Ten-year cross-sectional and incidence study of coronal and root caries and some related factors in elderly Swedish individuals. Gerodontology 2004;21(3):130–140. [PubMed: 15369015]
21. Lawrence HP, Hunt RJ, Beck JD. Three-year root caries incidence and risk modeling in older adults in North Carolina. Journal of public health dentistry 1995;55(2):69–78. [PubMed: 7643330]
22. McDermott RE, Hoover JN, Komiyama K. Root surface caries prevalence and associated factors among adult patients in an acute care hospital. J Can Dent Assoc 1991;57(6):505–508. [PubMed: 1860090]
23. Tan HP, Lo ECM. Risk indicators for root caries in institutionalized elders. Community Dentistry and Oral Epidemiology 2014;42(5):435–440. [PubMed: 24750310]
24. RA V, SD A, BJ D. Root caries risk indicators: a systematic review of risk models. Community Dentistry and Oral Epidemiology 2010;38(5):383–397. [PubMed: 20545716]
25. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology 2013;5(5):241–266.
26. Jothi N, Husain W. Data mining in healthcare–a review. Procedia Computer Science 2015;72:306–313.
27. National Health and Nutrition Examination Survey Data In. CfDCaP, trans. U.S. Department of Health and Human Services Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS) https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015. . Hyattsville, MD2015-2016.
28. Marcenes W, Kassebaum NJ, Bernabe E, et al. Global burden of oral conditions in 1990–2010: a systematic analysis. Journal of dental research 2013;92(7):592–597. [PubMed: 23720570]
29. Griffin SO, Griffin PM, Swann JL, Zlobin N. Estimating rates of new root caries in older adults. Journal of dental research 2004;83(8):634–638. [PubMed: 15271973]
30. Banerjee P, Dehnbostel FO, Preissner R. Prediction Is a Balancing Act: Importance of Sampling Methods to Balance Sensitivity and Specificity of Predictive Models Based on Imbalanced Chemical Data Sets. Frontiers in chemistry 2018;6:362. [PubMed: 30271769]
31. Nath A, Subbiah K. Probing an optimal class distribution for enhancing prediction and feature characterization of plant virus-encoded RNA-silencing suppressors. 3 Biotech 2016;6(1):93.
32. Breiman L Random forests. Machine learning 2001;45(1):5–32.

33. Calle ML, Urrea V, Boulesteix AL, Malats N. AUC-RF: a new strategy for genomic profiling with random forest. Human heredity 2011;72(2):121–132. [PubMed: 21996641]
34. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. The Annals of Applied Statistics 2010;4(1):266–298.
35. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal 2015;13:8–17. [PubMed: 25750696]
36. Ritter AV, Shugars DA, Bader JD. Root caries risk indicators: a systematic review of risk models. Community Dent Oral Epidemiol 2010;38(5):383–397. [PubMed: 20545716]
37. U.S. Public Health Service DoHaHS. Oral Health in America: A report of the Surgeon General National Institute of Dental and Craniofacial Research2000.
38. Thomson WM, Williams SM, Broadbent JM, Poulton R, Locker D. Long-term Dental Visiting Patterns and Adult Oral Health. Journal of Dental Research 2010;89(3):307–311. [PubMed: 20093674]
39. Newman JF, Gift HC. Regular pattern of preventive dental services—A measure of access. Social Science & Medicine 1992;35(8):997–1001. [PubMed: 1411707]
40. Jansson L Association between alcohol consumption and dental health. Journal of Clinical Periodontology 2008;35(5):379–384. [PubMed: 18341603]
41. Sambunjak D. Flossing for the management of periodontal diseases and dental caries in adults.
42. Balhaddad A The Effect of Flossing on Dental Caries: A Critique of Current Literature. JOJ Case Stud 2017;4.
43. Estrela C, Guedes OA, Silva JA, Leles CR, Estrela CRdA, Pécora JD. Diagnostic and clinical factors associated with pulpal and periapical pain. Brazilian dental journal 2011;22(4):306–311. [PubMed: 21861030]
44. Chin LS-H, Chan JC-Y. Self-esteem, oral health behaviours, and clinical oral health status in Chinese adults: An exploratorystudy. Health Education Journal 2012;72(6):684–694.
45. Kojima A, Ekuni D, Mizutani S, et al. Relationships between self-rated oral health, subjective symptoms, oral health behavior and clinical conditions in Japanese university students: a cross-sectional survey at Okayama University. BMC oral health 2013;13(1):62. [PubMed: 24195632]
46. Mattila KJ, Nieminen MS, Valtonen VV, et al. Association between dental health and acute myocardial infarction. BMJ (Clinical research ed) 1989;298(6676):779–781.
47. Mattila KJ, Pussinen PJ, Paju S. Dental infections and cardiovascular diseases: a review. Journal of periodontology 2005;76(11 Suppl):2085–2088.
48. Gati D, Vieira AR. Elderly at greater risk for root caries: a look at the multifactorial risks with emphasis on genetics susceptibility. Int J Dent 2011;2011:647168. [PubMed: 21754932]
49. Older People Projected to Outnumber Children. 2017 National Population Projections 2018

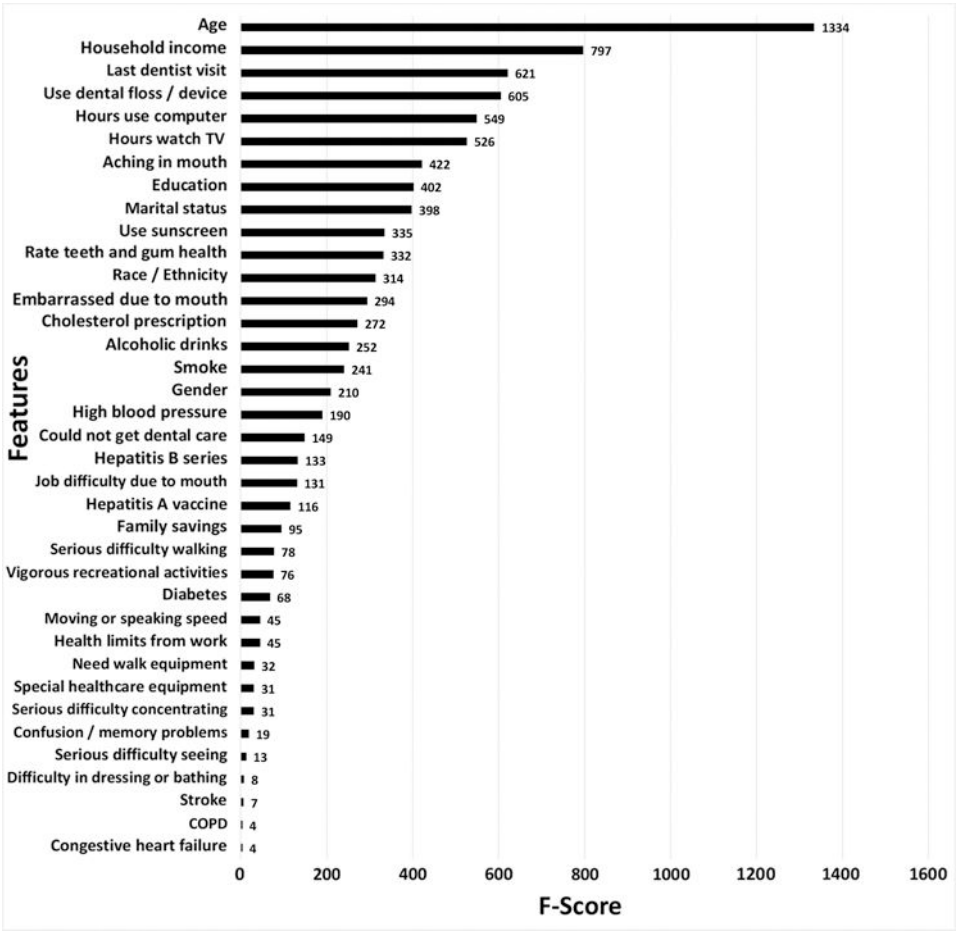

Figure 1.
Variable importance

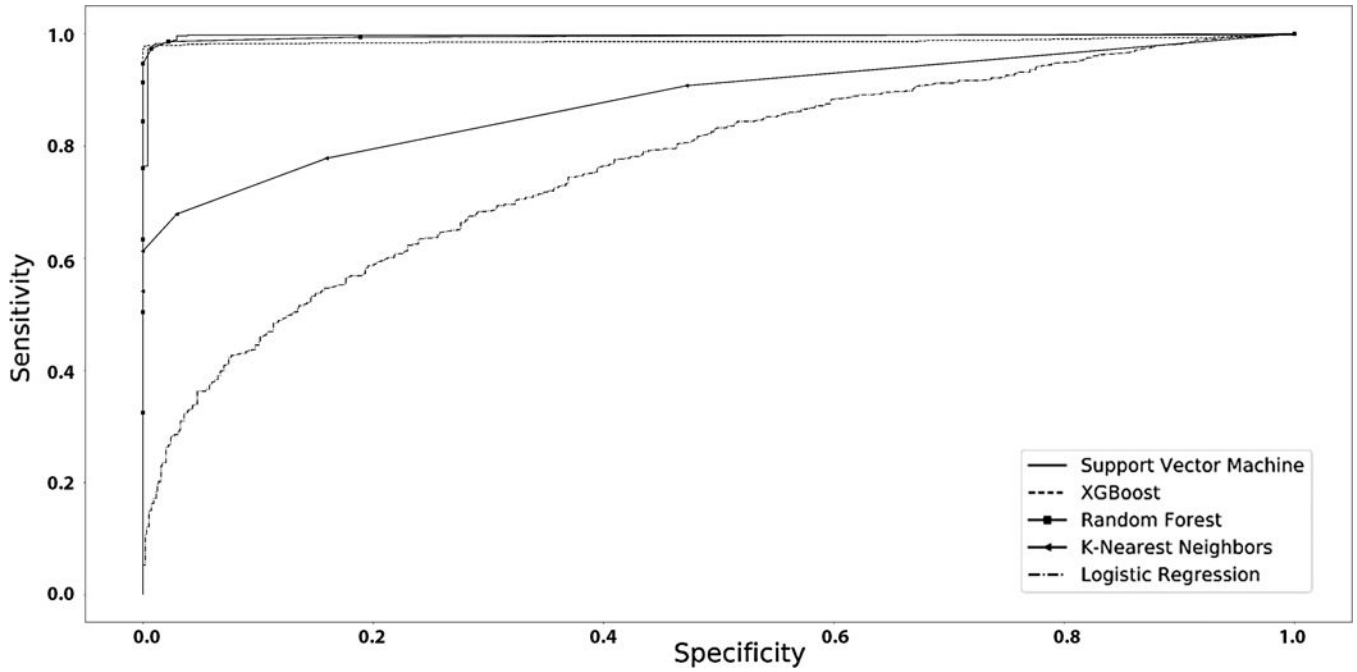

Figure 2.
ROC curves of the machine learning algorithms

Table 1:

Demographic characteristics (N = 5,135)

Variables	Mean	SD	Median	n	%
Age (year)	46.64	18.12	46.00	5,135	
Gender					
Male				2,485	48.4
Female				2,650	51.6
Race/Ethnicity					
Mexican American				940	18.3
Other Hispanic				673	13.1
Non-Hispanic White				1,629	31.7
Non-Hispanic Black				1,094	21.3
Non-Hispanic Asian				611	11.9
Other Race - Including Multi-Racial				188	3.7
Marital Status					
Married				2,510	48.9
Widowed				300	5.8
Divorced				479	9.3
Separated				162	3.2
Never married				930	18.1
Living with partner				493	9.6
Missing				261	5.1
Education level – xAdults 20+					
Less than 9^{th} grade				532	10.4
9^{th} – 11^{th} grade				544	10.6
High school graduate/GED or equivalent				1,051	20.5
Some college or AA degree				1,473	28.7
College graduate or above				1,273	24.8
Missing				262	5.1
Annual household income					
$0 to $4,999				117	2.3
$5,000 to $9,999				167	3.3
$10,000 to $14,999				274	5.3
$15,000 to $19,999				300	5.8
$20,000 to $24,999				302	5.9
$25,000 to $34,999				548	10.7
$35,000 to $44,999				487	9.5
$45,000 to $54,999				425	8.3
$55,000 to $64,999				322	6.3
$65,000 to $74,999				267	5.2
$20,000 and over				175	3.4
Under $20,000				76	1.5

Variables	Mean	SD	Median	n	%
$75,000 to $99,999				490	9.5
$100,000 and over				851	16.6
Missing				334	6.5

Table 2:

Top features selected in the machine learning algorithms

Variable Name	Variable Description	Total (N)	Root Caries		p-value*
			Y(n)	N(n)	
PAQ710	Hours watch TV or videos past 30 days	5,124	788	4,336	<.001
PAQ715	Hours use computer past 30 days	5,133	789	4,344	<.001
INDHHIN2	Annual Household Income	4,801	732	4,069	<.001
OHQ030	When did you last visit a dentist	5,122	789	4,333	<.001
DEQ034D	Use sunscreen?	3,397	439	2,958	<.001
OHQ845	Rate the health of your teeth and gums	5,133	791	4,342	<.001
DMDEDUC2	Education Level - Adults 20+	4,873	789	4,084	<.001
DMDMARTL	Marital Status	4,874	790	4,084	<.001
RIDRETH1	Race/Ethnicity	5,135	791	4,344	<.001
OHQ620	How often last year had aching in mouth?	3,962	731	3,231	<.001
OHQ680	Last year embarrassed because of mouth	3,966	731	3,235	<.001
ALQ151	Ever have 4/5 or more drinks every day?	3,865	620	3,245	<.001
BPQ090D	Told to take prescription for cholesterol	3,828	592	3,236	<.001
OHQ870	How many days use dental floss/device	3,962	730	3,232	<.001
RIDAGEYR	Age in years at screening	5,135	791	4,344	<.001

Note:

* Independent samples t-tests for continuous variables, and chi-square tests for categorical variables.

Table 3:

Performance metrics of machine learning models using the top 15 selected features

Classifier	Accuracy	Precision	Sensitivity	Specificity	AUC
Support Vector Machine	0.971	0.951	0.996	0.943	0.997
XGBoost	0.947	0.908	1.000	0.889	0.987
Random Forest	0.941	0.947	1.000	0.875	0.999
k-Nearest Neighbors	0.832	0.769	0.971	0.679	0.881
Logistic Regression	0.742	0.742	0.771	0.711	0.818