Original Article

Identifying Factors Associated with Periodontal Disease Using Machine Learning

Hussam M. Alqahtani[1,2,3], Siran M. Koroukian[1], Kurt Stange[1,4], Nicholas K. Schiltz[1], Nabil F. Bissada[5]

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA, [2]Department of Preventive Dental Science, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia, [3]King Abdullah International Medical Research Center, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia, [4]Center for Community Health Integration, Case Comprehensive Cancer Center, Cleveland, OH, USA, [5]Department of Periodontics, Case Western Reserve University School of Dental Medicine, Cleveland, OH, USA

Received : 17-09-22
Revised : 20-11-22
Accepted : 24-11-22
Published : 30-12-22

ABSTRACT

Objective: This study aimed to identify combinations of chronic conditions associated with the presence and severity of periodontal disease (PD) after accounting for a series of demographic and behavioral characteristics in a nationally representative sample of US adults. **Materials and Methods:** A cross-sectional study of the 2013–2014 National Health and Nutrition Examination Survey (n = 4555). Outcome measure: PD using clinical attachment loss (measured as none, mild, moderate, or severe). The main independent variables were self-reported chronic conditions, while other covariates included demographic and behavioral variables. Classification and regression tree analysis was used to identify combinations of specific chronic conditions associated with PD and PD with higher severity. Random forest was used to identify the most important variables associated with the presence and severity of PD. **Results:** The prevalence of PD was 77% among the study population. The percentage of those with PD was higher among younger and middle-aged (< 61 years old) than older (> 61 years old) adults. Age and education level were the two most important predictors for the presence and severity of PD. Other significant factors included alcohol use, type of medical insurance, sex, and non-white race. Accounting for only chronic conditions, hypertension and diabetes were the two chronic conditions associated with the presence and severity of PD. **Conclusions:** Sociodemographic and behavioral factors emerged as more strongly associated with the presence and severity of PD than chronic conditions. Accounting for the co-occurrence for sociodemographic and behavioral factors will be informative for identifying people vulnerable to the development of PD.

Keywords: *Machine learning, periodontal medicine, periodontal-systemic disease interactions, periodontitis, risk factor(s)*

Introduction

Periodontal disease (PD) is a chronic multifactorial inflammatory disease. It is associated with dysbiosis plaque biofilms resulting in chronic destructive inflammatory responses.[1] In its mildest form, it affects 45%-50% of adults.[2] Severe PD is the sixth most common disease, and it is estimated to affect 11.2% of the global adult population.[2]

Several risk factors and indicators are shared between PD and several systemic conditions, including cardiovascular disease, obesity, rheumatoid arthritis, prostate cancer, stroke, cognitive impairment, and hypertension.[3] Suggested mechanisms include bacteremia through the epithelium lining of periodontal pockets and elevation of systemic inflammatory cytokines.[4]

Access this article online

Quick Response Code:

Website: www.jispcd.org

DOI: 10.4103/jispcd.JISPCD_188_22

Address for correspondence: Dr. Hussam M. Alqahtani, Department of Preventive Dental Science, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia.
E-mail: ude.esac@53amh

How to cite this article: Alqahtani HM, Koroukian SM, Stange K, Schiltz NK, Bissada NF. Identifying factors associated with periodontal disease using machine learning. J Int Soc Prevent Communit Dent 2022;12:612-20.

Prior studies have confirmed the relationship between systemic conditions and PD.[3-6] However, the growing presence of people living with multiple chronic conditions is likely to shift priorities in health care management toward conditions affecting systemic health, placing PD care at a lower priority.[7-9] Consequently, multiple chronic conditions are likely to impact their periodontal condition negatively. Assessing the relationship between PD and a given chronic condition does not account for common co-occurring chronic conditions, underestimating the compounding effects of other chronic conditions.[10] Therefore, we hypothesize that there are combinations of specific chronic conditions associated with the presence of PD and PD with higher severity, accounting for a series of demographic and behavioral characteristics.

To the best of our knowledge, no previous studies have identified co-occurring chronic conditions associated with PD or its severity. Identifying the most prevalent combinations of chronic conditions that are associated with the presence and severity of PD will not only elucidate etiologic and pathophysiologic pathways but will also help to raise awareness among healthcare providers regarding PD, and prompt them to recommend periodic periodontal checkups for people who are most vulnerable to the development or worsening of PD.

This study aimed to identify combinations of chronic conditions that are associated with the presence and severity of the PD, after accounting for a series of demographic and behavioral characteristics.

Materials and Methods

This cross-sectional study uses the publicly available data from the 2013–2014 National Health and Nutrition Examination Survey (NHANES), the most recent years with data on PD, using full-mouth periodontal examination protocol. NHANES is a cross-sectional survey intended to observe the overall health and nutritional status of a nationally representative sample of the U.S population. This study was deemed research not involving human subjects by the Case Western Reserve University Institutional Review Board (# 2021-0469)

Data source and study population

NHANES conducts health interviews as well as examinations ranging from laboratory tests to physiological measurements. NHANES interviews and examinations are performed by trained medical personnel, and the information collected for the NHANES surveys is done via a multistage probability design. Our study population included 4,669 individuals.

Variables of interest

Outcome variable

Our outcome variable was PD (present/absent), using clinical attachment loss (CAL) according to the 2018 American Academy of Periodontology (AAP) classification.[11] Those categorized as PD were further grouped in the category of mild (CAL=1–2 mm), moderate (CAL=3–4 mm), or severe PD (CAL > 5 mm).

Independent variables

Self-reported chronic conditions indicating whether a physician ever told the individual that he or she had, including: hypertension, hyperlipidemia, diabetes, arthritis, coronary heart disease, overweight, stroke, asthma, chronic obstructive pulmonary disease, emphysema, chronic bronchitis, cancer, liver condition, thyroid problems, psoriasis and weak or failing kidneys.

Other variables of interest included: age (<30, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, >80), sex (Male, Female), race/ethnicity (Hispanic, White, Black, Other), marital status (married, widowed, divorced, never married), education (< 9th grade, 9-12th grade, high school graduate or equivalent, some college or associates degree, College graduate or higher), the ratio of family income to poverty (<1, 1–1.99, 2–2.99, 3–3.99, 4–4.99, >5), smoking status (Current, Former, None), alcohol use (Yes, No), body mass index (underweight (<18 kg/m^2), normal/ overweight (18.1–30 kg/m^2), obese (>30 kg/m^2)), vigorous recreational activities (Yes, No), insurance status (Medicare, Medicaid, Private insurance, and All other) and whether their insurance plan provides coverage for dental procedures (Yes, No). For variables with missing values of more than 1% (insurance status, ratio of family income to poverty, and body mass index), we created a missing category. We excluded observations when missing values amounted to less than 1% missing. The three variables that fit that category were education, marital status, and smoking behavior. Our final sample size includes 4555 participants after excluding 114 participants with missing data in our variables of interest.

Statistical analysis

We began our analysis by conducting a descriptive analysis of our study variables. Regarding our outcome variable, we first examined correlates of PD as present/absent; and among those with PD, we identified factors associated with moderate/severe PD. Next, we identified the most common combinations of chronic conditions associated with the presence and severity of PD using a

conditional inference regression tree (CTree), described below.[12] Although CTree is similar to the classification and regression tree (CART),[13] it uses a statistical significance test as the splitting criterion. We addressed our aim in two steps. First, we identified which combinations of chronic conditions, sociodemographic and behavioral variables are associated with PD. Subsequently, limiting our analysis to respondents with PD, we will identify the most common combinations of chronic conditions most highly associated with moderate/severe PD.

CTree, a machine learning method, is a recursive binary partitioning of the data with each variable's ability to be considered a potential split. Each node can split and form two child nodes, which can successively split and create two more child nodes. This process continues until a node can no longer be split given the stopping criteria, thereby creating a terminal node. The following stopping criteria (a maximum tree depth of five splits, a minimum terminal node size of 100 participants, and a p-value threshold of 0.001) were used. For the CTree analysis of PD severity using only chronic conditions with a minimum of 500 positive cases, i.e., hypertension, hyperlipidemia, diabetes, arthritis, asthma, bronchitis, cancer, liver condition, and thyroid problems, 0.001 might be too strict given the smaller sample size of the PD subpopulation; therefore, we used an alpha of 0.05.

We built the CTree model by partitioning the data into training and test datasets. We used the validated dataset to test the accuracy of the CTree model. In addition, we used a random forest approach to select the best predictors to partition the outcome at each node. Random forest is a bootstrap aggregation method that builds a tree using a random variable selection.[14] We created 3,000 trees and sampled three of the explanatory variables at each node split for each random forest model. Next, we compared the two models, CTree, and Random forest, to check the agreement on the most important predictors for PD and moderate/severe PD. We used R version 3.6 and the "partykit" (CTree), "randomForest" (random forest) packages.

Results

The characteristics of the study population by presence and severity of PD are presented in [Table 1]. Of the 4555 participants, 77.4% had PD. Among those with PD, 27.0% presented with mild PD, 54.3% with moderate PD, and 18.7% with severe PD. The percentage of those with PD was higher among middle-aged (two-thirds) than older individuals. There were more men in the severe PD group. Across all PD groups, a higher proportion of participants were married, non-smokers, normal/overweight and obese, and those without vigorous physical activity. Nearly 30% of study participants with severe PD had missing values on their type of health insurance. About 96% of the study population across all PD groups did not have dental insurance.

Findings from [Table 2] showed that hypertension, hyperlipidemia, diabetes, arthritis, and overweight were the most common chronic conditions across all PD groups. Among those without PD, coronary heart disease, stroke, chronic obstructive pulmonary disease, and emphysema were the most common chronic conditions. Out of all chronic conditions, hypertension, hyperlipidemia, diabetes, arthritis, overweight, and asthma had a minimum of 500 positive cases of PD.

[Figure 1] shows the CTree analysis for the presence and severity of PD using chronic conditions, sociodemographic and behavioral variables. The tree shows the different distribution of PD with different combinations of sociodemographic and behavioral variables. The highest percentage of PD was observed among participants who had a non-missing value for alcohol consumption, were privately insured, or had missing values on type of insurance and were 36 years of age or less (nodes 1, 2, 3, 4, and 5), leading to over 90% of respondents with this combination of sociodemographic and behavioral variables reported PD. Conversely, the lowest prevalence of PD was observed among participants who were 61 years of age or younger and had missing values on alcoholic consumption (nodes 1, 2, and 8), leading to about 55% prevalence of PD. The highest percentage of moderate/severe PD (about 90%) was observed among those who had high school graduate education levels or less and were male (nodes 1, 2, and 3). When education levels were some college degrees or higher, co-occurred with an age of 50 years or younger, white race placed them at low risk of moderate/severe PD at nearly 45% (nodes 1, 5, 6, and 7) compared to other types of races.

To address our primary research question, we subsequently limited our CTree analysis for the presence and severity of PD to chronic conditions only [Figure 2]. [Figure 2] shows the highest percentage of PD included individuals with no hypertension, no arthritis, and no diabetes (nodes 1, 5, 7, and 9), leading to over 80% of participants with a combination of conditions reported PD. Although the prevalence of PD among participants with different combinations of the chronic condition is generally high, the lowest prevalence of PD was observed among participants with blood pressure and arthritis (nodes 1, 2, and 3), leading to a prevalence of PD slightly higher than 60%. In addition, [Figure 2]

Table 1: Demographic characteristics of the study population

Demographic Characteristics	No PD	Mild PD	Moderate PD	Severe PD	Total (PD+no PD)	Total PD
No. of subjects	1030	954	1913	658	4555	3525
Age categories						
[30-35]	47 (9.5)	155 (31.2)	262 (52.7)	33 (6.6)	497	450 (90.5)
[35-40]	54 (11.8)	157 (34.4)	209 (45.8)	36 (7.9)	456	402 (88.2)
[40-45]	77 (14.7)	159 (30.3)	243 (46.4)	45 (8.6)	524	447 (85.3)
[45-50]	78 (16.8)	133 (28.7)	189 (40.7)	64 (13.8)	464	386 (83.2)
[50-55]	87 (19.1)	94 (20.7)	192 (42.2)	82 (18.0)	455	368 (80.9)
[55-60]	85 (19.2)	71 (16.0)	199 (44.9)	88 (19.9)	443	358 (80.8)
[60-65]	137 (27.5)	62 (12.4)	182 (36.5)	118 (23.6)	499	362 (72.5)
[65-70]	114 (29.5)	52 (13.4)	141 (36.4)	80 (20.7)	387	273 (70.5)
[70-75]	124 (38.4)	35 (10.8)	123 (38.1)	41 (12.7)	323	199 (61.6)
[75-80]	84 (44.0)	12 (6.3)	67 (35.1)	28 (14.7)	191	107 (56.0)
[80-81]	143 (45.3)	24 (7.6)	106 (33.5)	43 (13.6)	316	173 (54.7)
Sex						
Male	458 (21.2)	339 (15.7)	937 (43.4)	426 (19.7)	2160	1702 (78.8)
Female	572 (23.9)	615 (25.7)	976 (40.8)	232 (9.7)	2395	1823 (76.1)
Race/Ethnicity						
White	478 (23.9)	530 (26.6)	788 (39.5)	200 (10.0)	1996	1518 (76.1)
Black	233 (25.2)	128 (13.8)	369 (39.8)	196 (21.2)	926	693 (74.8)
Hispanic	201 (20.4)	152 (15.4)	455 (46.1)	178 (18.1)	986	785 (79.6)
Other	118 (18.2)	144 (22.3)	301 (46.5)	84 (13.0)	647	529 (81.8)
Marital status						
Married	552 (19.1)	686 (23.8)	1239 (43.0)	406 (14.1)	2883	2331 (80.9)
Divorced	189 (24.9)	149 (19.6)	304 (40.1)	117 (15.4)	759	570 (75.1)
Widowed	172 (42.9)	32 (8.0)	141 (35.2)	56 (14.0)	401	229 (57.1)
Never married	117 (22.9)	87 (17.0)	229 (44.7)	79 (15.4)	512	395 (77.1)
Education						
Less than 9th grade	139 (34.2)	22 (5.4)	156 (38.4)	89 (21.9)	406	267 (65.8)
9-12th grade	190 (31.7)	65 (10.8)	212 (35.3)	133 (22.2)	600	410 (68.3)
High school graduate or equivalent	259 (25.3)	160 (15.6)	406 (39.6)	199 (19.4)	1024	765 (74.7)
Some college or associates degree	253 (19.2)	302 (22.9)	608 (46.1)	155 (11.8)	1318	1065 (80.8)
College graduate or above	189 (15.7)	405 (33.6)	531 (44.0)	82 (6.8)	1207	1018 (84.3)
Ratio of family income to poverty						
[0,1]	254 (29.6)	96 (11.2)	320 (37.3)	187 (21.8)	857	603 (70.4)
[1,2]	289 (26.7)	164 (15.2)	448 (41.4)	180 (16.7)	1081	792 (73.3)
[2,3]	115 (20.6)	112 (20.1)	249 (44.6)	82 (14.7)	558	443 (79.4)
[3,4]	102 (18.7)	135 (24.8)	238 (43.7)	70 (12.8)	545	443 (81.3)
[4,5]	57 (18.2)	90 (28.7)	139 (44.3)	28 (8.9)	314	257 (81.8)
[5,6]	125 (14.8)	299 (35.5)	361 (42.9)	57 (6.8)	842	717 (85.2)
Missing	88 (24.6)	58 (16.2)	158 (44.1)	54 (15.1)	358	270 (75.4)
Smoking status						
No	484 (19.5)	639 (25.7)	1102 (44.3)	261 (10.5)	2486	2002 (80.5)
Former	294 (25.1)	211 (18.0)	476 (40.6)	191 (16.3)	1172	878 (74.9)
Current	252 (28.1)	104 (11.6)	335 (37.3)	206 (23.0)	897	645 (71.9)
Alcohol use						
Yes	550 (18.6)	655 (22.1)	1298 (43.8)	460 (15.5)	2963	2413 (81.4)
No	283 (23.9)	238 (20.1)	514 (43.4)	150 (12.7)	1185	902 (76.1)
Missing	197 (48.4)	61 (15.0)	101 (24.8)	48 (11.8)	407	210 (51.6)
Body mass index						
Missing	40 (67.8)	*	*	*	59	19 (32.2)
Underweight	19 (44.2)	*	14 (32.6)	*	43	24 (55.8)
Normal/overweight	581 (21.4)	624 (23.0)	1118 (41.1)	395 (14.5)	2718	2137 (78.6)
Obese	390 (22.5)	325 (18.7)	772 (44.5)	248 (14.3)	1735	1345 (77.5)

Table 1: Demographic characteristics of the study population

Demographic Characteristics	No PD	Mild PD	Moderate PD	Severe PD	Total (PD+no PD)	Total PD
Vigorous exercise						
No	935 (25.2)	677 (18.2)	1511 (40.7)	591 (15.9)	3714	2779 (88.7)
Yes	95 (11.3)	277 (32.9)	402 (47.8)	67 (8.0)	841	746 (74.8)
Insurance						
Missing	155 (17.7)	128 (14.6)	401 (45.7)	194 (22.1)	878	723 (82.3)
Medicaid	163 (34.2)	56 (11.8)	164 (34.5)	93 (19.5)	476	313 (65.8)
Medicare	337 (38.4)	93 (10.6)	330 (37.6)	117 (13.3)	877	540 (61.6)
Other	110 (26.2)	78 (18.6)	162 (38.6)	70 (16.7)	420	310 (73.8)
Private	265 (13.9)	599 (31.5)	856 (45.0)	184 (9.7)	1904	1639 (86.1)
Insurance cover the dental procedure						
No	995 (22.5)	932 (21.1)	1865 (42.2)	630 (14.2)	4422	3427 (77.5)
Yes	35 (26.3)	22 (16.5)	48 (36.1)	28 (21.1)	133	98 (73.7)

* Cells with counts < 11 were masked

Table 2: Description of the study population by chronic conditions

Chronic Conditions	No PD	Mild PD	Moderate PD	Severe PD	Total (PD+no PD)	Total PD
No. of subjects	1030	954	1913	658	4555	3525
Hypertension	574 (29.8)	304 (15.8)	739 (38.4)	308 (16.0)	1925	1351 (70.2)
Hyperlipidemia	475 (25.7)	370 (20.0)	763 (41.3)	241 (13.0)	1849	1374 (74.3)
Diabetes	260 (31.6)	99 (12.0)	314 (38.2)	149 (18.1)	822	562 (68.4)
Arthritis	421 (30.2)	232 (16.6)	546 (39.2)	195 (14.0)	1394	973 (69.8)
Coronary heart disease	106 (50.0)	18 (8.5)	62 (29.2)	26 (12.3)	212	106 (50.0)
Overweight	372 (22.0)	341 (20.2)	744 (44.0)	232 (13.7)	1689	1317 (78.0)
Stroke	86 (48.6)	18 (10.2)	43 (24.3)	30 (16.9)	177	91 (51.4)
Asthma	166 (24.7)	162 (24.1)	280 (41.6)	65 (9.7)	673	507 (75.3)
Chronic obstructive pulmonary disease	82 (46.3)	13 (7.3)	52 (29.4)	30 (16.9)	177	95 (53.7)
Emphysema	48 (56.5)	*	17 (20.0)	14 (16.5)	85	37 (43.5)
Bronchitis	90 (33.1)	31 (11.4)	112 (41.2)	39 (14.3)	272	182 (66.9)
Cancer	160 (31.8)	78 (15.5)	196 (39.0)	69 (13.7)	503	343 (68.2)
Liver	60 (28.0)	35 (16.4)	91 (42.5)	28 (13.1)	214	154 (72.0)
Thyroid	159 (29.3)	122 (22.5)	221 (40.8)	40 (7.4)	542	383 (70.7)
Psoriasis	38 (29.0)	22 (16.8)	61 (46.6)	*	131	93 (71.0)
Kidney	57 (35.4)	20 (12.4)	58 (36.0)	26 (16.1)	161	104 (64.6)

* Cells with counts < 11 were masked

shows that the highest percentage of moderate/severe PD (over 80%) was observed among those with diabetes (nodes 1 and 2). In contrast, the lowest percentage of moderate/severe PD (about 60%) was observed among those who had no diabetes but had hypertension and asthma (nodes 1, 3, 4, and 5, leading to the second bar at the left-hand side). The prevalence of moderate/severe PD ranged between sixty and slightly higher than eighty percent among participants with different combinations of the chronic condition.

Random Forest analysis shows which variables are the most important in improving the model's accuracy from a bootstrap sample of 3000 trees for each outcome [Figures 3 and 4]. Using chronic conditions, sociodemographic and behavioral variables in [Figure 3], the following variables rank in the top three most important variables for PD: Age, alcohol use, and health insurance. For moderate/severe PD, age, education level, type of health insurance, the ratio of family income to poverty, gender, and race are the top six most important variables using chronic conditions, sociodemographic and behavioral variables. The frequent appearance of these variables in the CTree models in [Figure 1] validates those models and demonstrates the importance of sociodemographic and behavioral

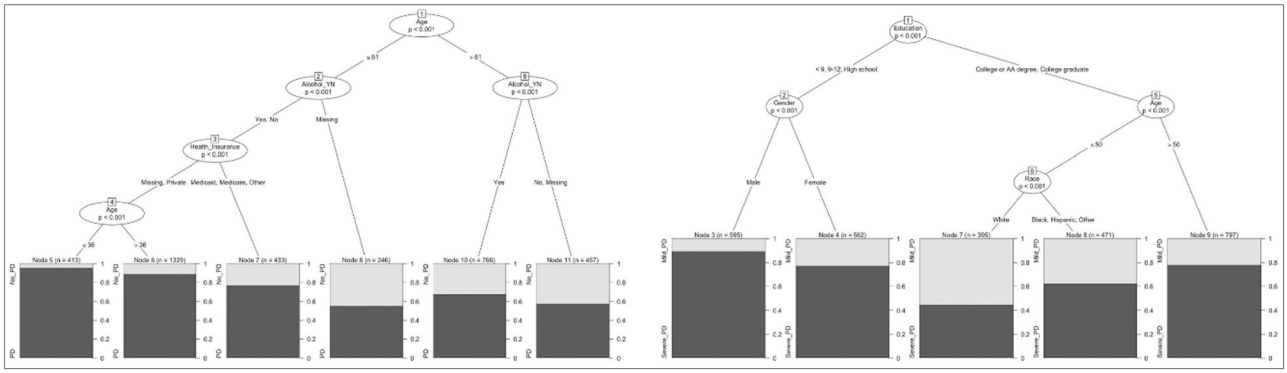

Figure 1: Conditional inference regression tree analysis to predict the presence of PD (left) and moderate/severe PD among those with PD (right). PD, Periodontitis; No_PD, No Periodontitis; Mild_PD, Mild Periodontitis; Severe_PD, Severe Periodontitis; Alcohol_YN, Alcohol consumption (Yes, No, missing values)

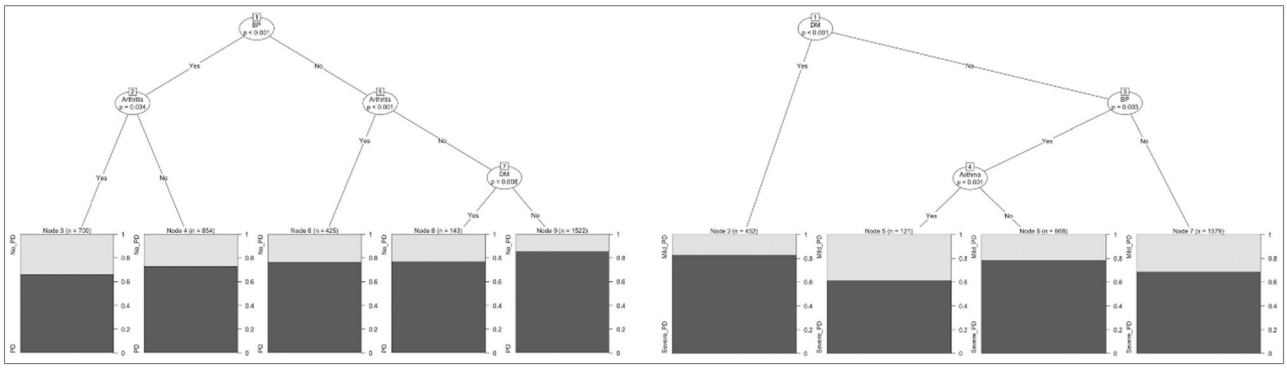

Figure 2: Conditional inference regression tree analysis to predict the presence of PD (left) and moderate/severe PD among those with PD (right) using only chronic conditions. PD, Periodontitis; No_PD, No Periodontitis; Mild_PD, Mild Periodontitis; Severe_PD, Severe Periodontitis; BP, Hypertension; DM, Diabetes Mellitus

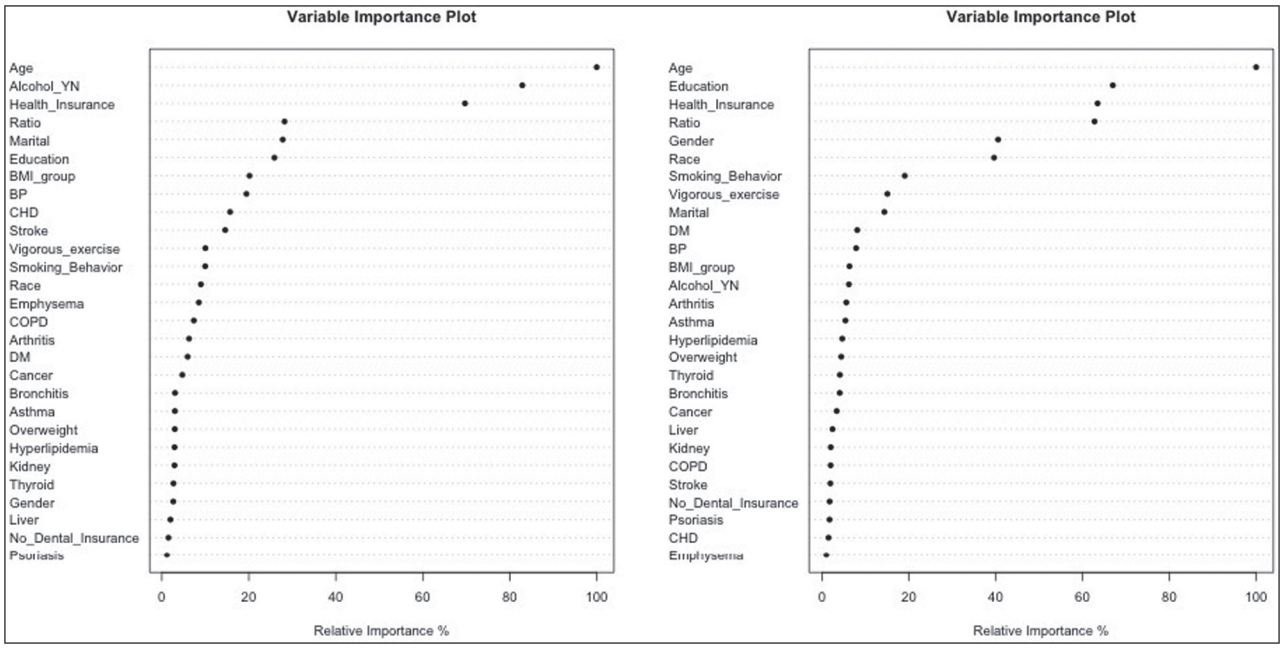

Figure 3: Random forest plot ranking the factors that most influence the distribution of PD (left plot) and moderate/severe PD (right plot). BP, Hypertension; CHD, Coronary Heart Disease; COPD, Chronic Obstructive Pulmonary Disease; DM, Diabetes; Alcohol_YN, Alcohol consumption

variables explaining the presence and severity of PD. We show logistic regression analysis for the presence of PD [Table 1S] and moderate/severe PD [Table 2S] in the supplemental material.

In [Figure 4], the Random Forest shows the top three most crucial variables for PD and moderate/severe PD using only chronic conditions. Hypertension, arthritis, and diabetes were the critical predictors for PD. For moderate/severe PD, diabetes, hypertension, and asthma are the most important variables. As most of these variables frequently appear in the CTree models in [Figure 2], they further validate those models. It also demonstrates the importance of hypertension and diabetes as factors explaining the presence and severity of PD, respectively.

Discussion

We analyzed the presence and severity of periodontal disease using a machine learning approach on data for a US representative sample of people from the NHANES. The higher prevalence of PD compared to previous estimates is likely due to the decrease in the CAL threshold with the new definition of PD.[11] In addition, the most recent case definition of PD used the CAL value on two non-adjacent teeth as primary criteria. In contrast, the old definition of PD was based on the CAL values and/or probing depth values on adjacent teeth. For comparison, we show the CTree analysis for the presence of PD, using old definition, in the supplemental material [Figure 1S]. Furthermore, the full-mouth periodontal examination protocol in the 2013–2014 NHANES allowed us to examine all teeth, which likely increases the accuracy of diagnosing and classifying PD compared with the use of partial-mouth periodontal examination protocols in earlier NHANES.[15]

Surprisingly, when we accounted for covariates in our CTree models, we showed that sociodemographic and behavioral variables were better predictors than chronic conditions for PD and PD severity. Age and education level were the most crucial variable for PD and moderate/severe PD, respectively. This can be explained by the high proportion of middle-aged adults in our study, who are less likely to have chronic conditions than older adults. In addition, the low prevalence of PD among older participants is likely because our sub-sample of NHANES for participants with periodontal data may be biased, representing a healthier subgroup of older people. For example, the 2013–2014 NHANES included only dentate participants; therefore, the edentate people who may have lost their teeth to PD, may not have participated in the oral health module of the NHANES, and were therefore not represented in our sample. In addition, dentate older people who are more likely to lose their teeth due to periodontal disease may have fewer teeth with better periodontal conditions, explaining the low prevalence of PD among older people. Moreover, the use of self-reported data may underestimate the true prevalence of chronic conditions due to participants forgetting or

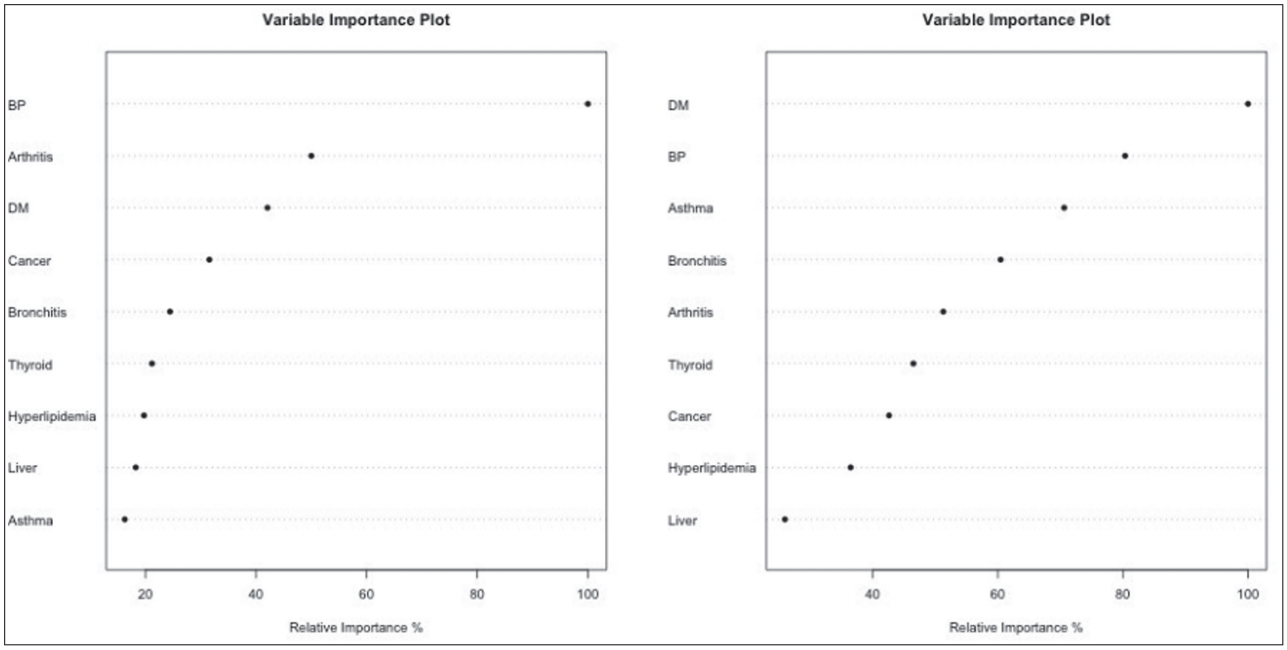

Figure 4: Random forest plot ranking the factors that most influence the distribution of PD (left plot) and moderate/severe PD (right plot), using only chronic conditions. BP, Hypertension; DM, Diabetes

not being diagnosed. Furthermore, about 45% of the study population had a low level of education, which may put their periodontal care at a lower priority. Other significant covariates included alcohol use, type of medical insurance, sex of participants, and participants' race. For example, two-thirds of the study population consumed alcohol compared to only one-fifth who were current smokers. This combination of sociodemographic and behavioral factors identified people who are at risk for the development of PD, given that PD is a prevalent disease that starts at an earlier age, regardless of the medical condition.

Given that chronic conditions did not appear in the models, we further developed CTree models including only chronic conditions to identify the characteristics of medically vulnerable people for the development of PD. In these models, hypertension and diabetes were the most important chronic conditions associated with the presence and severity of PD, respectively, as evidenced by the first splitting variable in the classification tree. Arthritis and asthma were the other critical chronic conditions that emerged from our machine learning approach.

Our results do not contradict the current knowledge on the impact of several chronic conditions on the development of PD. However, they highlight the sociodemographic and behavioral factors that may play a critical role in the development of PD at an earlier age and probably before the chronic conditions occurred. Except for a few chronic conditions, the prevalence of chronic conditions among study populations is very low; therefore, sociodemographic and behavioral factors have remarkably emerged among vulnerable people for PD regardless of their chronic condition.

These findings have important implications for clinical practice and research in the periodontal field. In clinical practice, identifying the most common combinations of sociodemographic and behavioral variables associated with the presence and severity of PD will elucidate etiologic, pathophysiologic, and behavioral pathways, beyond plaque biofilm. In addition, it will also help raise awareness among periodontists relative to the importance of including sociodemographic and behavior factors in their evaluation, specifically among middle-aged adults who are less prone to chronic conditions. Moreover, it prompts periodontists to recommend a periodic periodontal checkup for the most vulnerable people to the development or severity of PD earlier in age, regardless of chronic conditions. As a result, patient-centered care rather than disease-centered care is crucial in managing potential etiological factors involved in the development and severity of PD.

Regarding research implications, these findings indicate that accounting for sociodemographic and behavioral factors in addition to the co-occurrence for chronic conditions will be critical in evaluating PD. Research in this area had assessed the relationship between PD and a given chronic condition, but not accounted, for the common co-occurring chronic conditions, thus underestimating the compounding effects of other chronic conditions. As a result, the traditional focus on a single chronic condition limits our knowledge relative to the impact of co-occurring chronic conditions, sociodemographic and behavioral factors on PD.

To our knowledge, this is the first study to evaluate the patterns of co-occurrence of several factors that are associated with the presence and severity of PD using CTree, a novel machine learning method. The innovation of using this analytic approach allows us to discover the emerging combinations of the study factors without any prior hypotheses. CTree can capture the complex relationship and produce an easily interpretable decision tree model to identify specific combinations of the included factors highly associated with the presence and severity of PD. Random forest uses a subset of our data and bootstrapping to measure and rank the most important variable for our outcome. Although random forest cannot identify the most common combination of variables associated with PD, it can determine whether the top identified predictors agree with the most important variables that appear in CTree models. Therefore, both machine learning methods will detect the interaction and nonlinear relationship of our variables automatically. Another important strength of this study is the availability of several chronic conditions, and sociodemographic and behavioral factors in a nationally representative sample of the U.S population, allowing us to obtain deeper insight into the possible etiologies involved in developing PD.

There are several limitations of our study. First, with the use of cross-sectional data, it was not possible to capture worsening chronic conditions or PD. Second, all the chronic conditions are self-reported data, and only a small percentage of the population presented with chronic conditions. Third, CART produces a single tree, whereas Random Forest is a bootstrap aggregation method that produces multiple trees. However, many variables identified in CTree were also identified as the most important ones in Random Forest. Fourth, our models recognized that sociodemographic and

behavioral factors were stronger predictors for PD. However, for older people, for example, more expanded models including different types of variables such as tooth loss and oral hygiene may reflect an accurate estimate and identify additional predictors for PD and moderate/severe PD.

Conclusions

Sociodemographic and behavioral factors were better predictors for periodontal disease than chronic conditions. Hypertension and diabetes were the most critical chronic conditions that predict the presence and severity of the periodontal disease. Compared to chronic conditions, accounting for the co-occurrence of sociodemographic and behavioral factors is more informative when identifying people who are at heightened risk to develop PD.

Acknowledgements

None.

Financial support and sponsorship

None.

Conflicts of interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors report no conflicts of interest related to this study.

Author's contributions

Not applicable.

Ethical policy and institutional review board statement

This study was deemed research not involving human subjects by the Case Western Reserve University Institutional Review Board (#2021-0469).

Patient declaration of consent

Not applicable.

Data availability statement

Not applicable.

References

1. Papapanou PN, Sanz M, Buduneli N, Dietrich T, Feres M, Fine DH, *et al*. Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. J Clin Periodontol 2018;45(Suppl 20):S162-S170.
2. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the global burden of disease study 2013. Lancet 2015;386:743-800.
3. Hajishengallis G, Chavakis T. Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. Nat Rev Immunol 2021;21:426-40.
4. Hajishengallis G, Chavakis T, Lambris JD. Current understanding of periodontal disease pathogenesis and targets for host-modulation therapy. Periodontol 2000 2020;84:14-34.
5. Wu CZ, Yuan YH, Liu HH, Li SS, Zhang BW, Chen W, *et al*. Epidemiologic relationship between periodontitis and type 2 diabetes mellitus. BMC Oral Health 2020;20:204.
6. Sanz M, Marco Del Castillo A, Jepsen S, Gonzalez-Juanatey JR, D'Aiuto F, Bouchard P, *et al*. Periodontitis and cardiovascular diseases: Consensus report. J Clin Periodontol 2020;47:268-88.
7. Mariotti A, Hefti AF. Defining periodontal health. BMC Oral Health 2015;15 Suppl 1:S6.
8. Hajat C, Stein E. The global burden of multiple chronic conditions: A narrative review. Prev Med Rep 2018;12:284-93.
9. Boersma P, Black LI, Ward BW. Prevalence of multiple chronic conditions among US adults, 2018. Prev Chronic Dis 2020;17:E106.
10. Alqahtani HM, Koroukian SM, Stange K, Bissada NF, Schiltz NK. Combinations of chronic conditions, functional limitations and geriatric syndromes associated with periodontal disease. Fam Med Community Health 2022;10:e001733. doi: 10.1136/fmch-2022-001733. PMID: 35998996; PMCID: PMC9403150.
11. Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. J Periodontol 2018;89 Suppl 1:159-72.
12. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 2006;15:651-674.
13. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Routledge; 2017. doi: 10.1201/9781315139470.
14. Yoo W, Ference BA, Cote ML, Schwartz A. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions. Int J Appl Sci Technol 2012; 2:268.
15. Eke PI, Thornton-Evans GO, Wei L, Borgnakke WS, Dye BA, Genco RJ. Periodontitis in US adults: National health and nutrition examination survey 2009–2014. J Am Dent Assoc 2018;149:576-588.e6.

Supplementary Material

Table 1S. Logistic regression model predicting the presence of PD

Variable	Odds Ratio	95% CI	P-value
BPNo	1.26	(1.03, 1.53)	0.023
HyperlipidemiaNo	1.00	(0.83, 1.21)	0.996
DMNo	1.17	(0.94, 1.47)	0.164
ArthritisNo	0.95	(0.78, 1.17)	0.647
CHDNo	1.73	(1.21, 2.47)	0.003
OverweightNo	0.81	(0.65, 1.02)	0.072
StrokeNo	1.67	(1.12, 2.49)	0.011
AsthmaNo	0.98	(0.76, 1.26)	0.864
COPDNo	1.21	(0.77, 1.90)	0.419
EmphysemaNo	1.78	(0.97, 3.28)	0.062
BronchitisNo	1.08	(0.75, 1.55)	0.687
CancerNo	1.09	(0.83, 1.42)	0.535
LiverNo	1.26	(0.86, 1.84)	0.239
ThyroidNo	1.10	(0.85, 1.43)	0.467
PsoriasisNo	1.32	(0.82, 2.12)	0.257
KidneyNo	1.15	(0.76, 1.75)	0.498
Age	0.96	(0.95, 0.97)	0.000
GenderFemale	0.92	(0.75, 1.11)	0.377
RaceBlack	0.94	(0.74, 1.19)	0.598
RaceHispanic	1.10	(0.85, 1.43)	0.449
RaceOther	1.09	(0.81, 1.48)	0.558
MaritalDivorced	0.85	(0.67, 1.09)	0.200
MaritalWidowed	0.85	(0.63, 1.14)	0.276
MaritalNever	0.74	(0.55, 0.99)	0.044
Education9-12	1.08	(0.77, 1.52)	0.663
EducationHigh school	1.33	(0.96, 1.86)	0.089
EducationCollege or AA degree	1.49	(1.06, 2.08)	0.021
EducationCollege graduate	1.46	(1.00, 2.12)	0.047
Ratio [1,2)	1.03	(0.80, 1.35)	0.796
Ratio [2,3)	1.35	(0.97, 1.90)	0.079
Ratio [3,4)	1.16	(0.81, 1.65)	0.422
Ratio [4,5)	1.06	(0.69, 1.63)	0.788
Ratio [5,6]	1.32	(0.92, 1.90)	0.132
RatioMissing	1.02	(0.72, 1.46)	0.902
Smoking_BehaviorFormer	1.01	(0.81, 1.26)	0.923
Smoking_BehaviorCurrent	0.66	(0.52, 0.84)	0.001
Alcohol_YNNo	0.92	(0.73, 1.15)	0.456
Alcohol_YNMissing	0.18	(0.14, 0.24)	0.000
BMI_groupUnderweight	3.15	(1.09, 9.10)	0.034
BMI_groupNormal/overweight	4.67	(2.28, 9.56)	0.000
BMI_groupObese	3.96	(1.93, 8.14)	0.000
Vigorous_exerciseNo	0.73	(0.56, 0.97)	0.032
Health_InsuranceMedicaid	0.74	(0.53, 1.03)	0.073
Health_InsuranceMedicare	0.89	(0.64, 1.23)	0.469
Health_InsuranceOther	0.84	(0.59, 1.20)	0.343
Health_InsurancePrivate	1.08	(0.82, 1.44)	0.576
No_Dental_Insurance1	0.97	(0.59, 1.59)	0.912

Table 2S. Logistic regression model predicting moderate/severe PD among those with PD

Variable	Odds Ratio	95% CI	P-value
BPNo	1.10	(0.88, 1.38)	0.39
HyperlipidemiaNo	1.40	(1.13, 1.72)	0.00
DMNo	0.74	(0.55, 1.01)	0.06
ArthritisNo	1.03	(0.80, 1.32)	0.82
CHDNo	1.05	(0.55, 2.03)	0.88
OverweightNo	0.93	(0.73, 1.18)	0.53
StrokeNo	1.11	(0.54, 2.29)	0.77
AsthmaNo	1.44	(1.11, 1.88)	0.01
COPDNo	0.55	(0.22, 1.38)	0.20
EmphysemaNo	1.73	(0.47, 6.37)	0.41
BronchitisNo	0.43	(0.25, 0.73)	0.00
CancerNo	0.79	(0.55, 1.14)	0.21
LiverNo	0.94	(0.58, 1.51)	0.80
ThyroidNo	1.12	(0.83, 1.50)	0.46
PsoriasisNo	0.77	(0.43, 1.37)	0.38
KidneyNo	1.17	(0.64, 2.15)	0.60
Age	1.04	(1.03, 1.06)	0.00
GenderFemale	0.42	(0.34, 0.51)	0.00
RaceBlack	2.17	(1.63, 2.87)	0.00
RaceHispanic	1.79	(1.37, 2.34)	0.00
RaceOther	2.08	(1.56, 2.78)	0.00
MaritalDivorced	0.88	(0.67, 1.14)	0.33
MaritalWidowed	1.72	(1.01, 2.94)	0.05
MaritalNever	1.33	(0.96, 1.84)	0.09
Education9-12	0.58	(0.31, 1.06)	0.08
EducationHigh school	0.47	(0.27, 0.84)	0.01
EducationCollege or AA degree	0.38	(0.22, 0.67)	0.00
EducationCollege graduate	0.31	(0.17, 0.56)	0.00
Ratio [1,2)	0.70	(0.50, 1.00)	0.05
Ratio [2,3)	0.73	(0.49, 1.09)	0.12
Ratio [3,4)	0.60	(0.40, 0.90)	0.01
Ratio [4,5)	0.59	(0.37, 0.93)	0.02
Ratio [5,6]	0.46	(0.31, 0.69)	0.00
RatioMissing	0.75	(0.47, 1.18)	0.22
Smoking_BehaviorFormer	1.01	(0.79, 1.28)	0.95
Smoking_BehaviorCurrent	2.02	(1.49, 2.72)	0.00
Alcohol_YNNo	0.99	(0.78, 1.25)	0.93
Alcohol_YNMissing	1.01	(0.67, 1.53)	0.95
BMI_groupUnderweight	0.84	(0.07, 10.12)	0.89
BMI_groupNormal/overweight	0.30	(0.04, 2.53)	0.27
BMI_groupObese	0.37	(0.04, 3.05)	0.36
Vigorous_exerciseNo	1.15	(0.91, 1.44)	0.24
Health_InsuranceMedicaid	0.71	(0.46, 1.10)	0.13
Health_InsuranceMedicare	0.70	(0.45, 1.09)	0.12
Health_InsuranceOther	0.67	(0.45, 1.01)	0.06
Health_InsurancePrivate	0.68	(0.51, 0.91)	0.01
No_Dental_Insurance1	0.85	(0.46, 1.56)	0.60

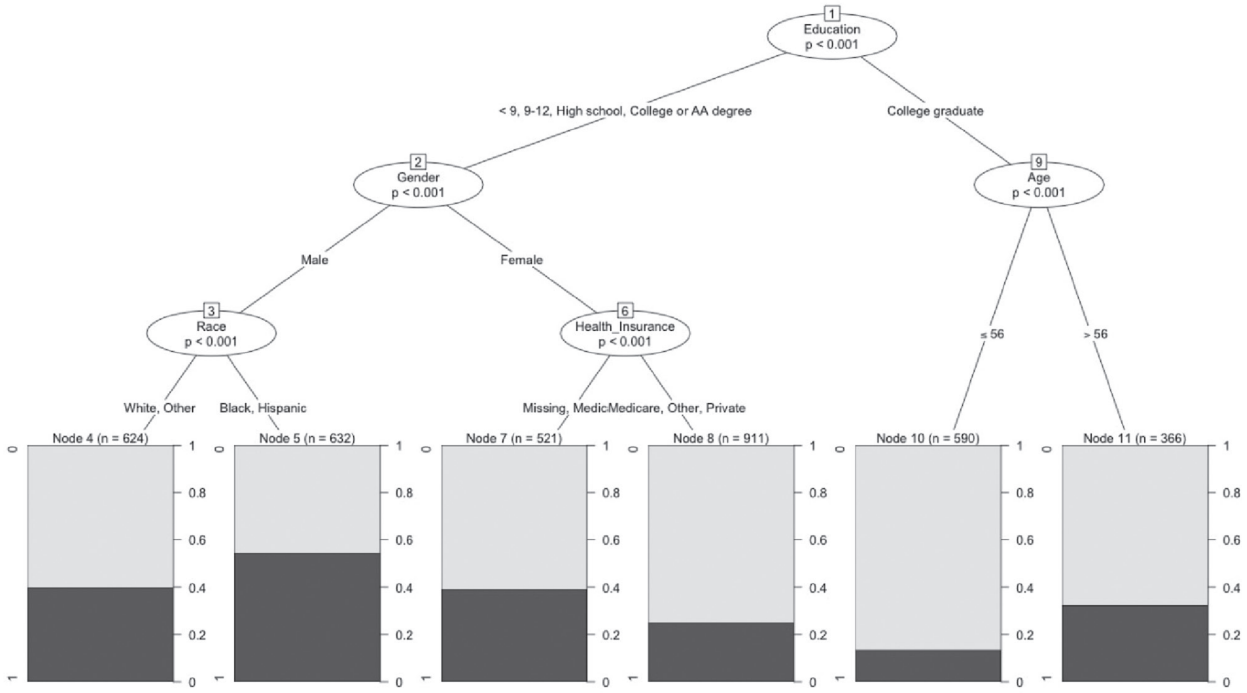

Figure 1S: Conditional inference regression tree analysis to predict the presence of PD using old definition. 0, No Periodontitis; 1, Periodontitis