Received: 23 July 2020 | Revised: 28 November 2020 | Accepted: 3 December 2020

DOI: 10.1002/JPER.20-0489

JOURNAL OF Periodontology

ORIGINAL ARTICLE

# Clustering by periodontitis-associated factors: A novel application to NHANES data

**Iya H. Ghassib**[1] | **Feras A. Batarseh**[2] | **Hom-Lay Wang**[1] | **Wenche S. Borgnakke**[1]

[1] School of Dentistry, Department of Periodontics and Oral Medicine, University of Michigan, Ann Arbor, MI

[2] College of Science, George Mason University, Fairfax, VA

**Correspondence**
Iya Ghassib DDS, Department of Periodontics and Oral Medicine, School of Dentistry, University of Michigan, 1011 North University Avenue, Ann Arbor, MI 48109-1078, USA.
Email: ighassib@umich.edu

## Abstract

**Background:** Unsupervised clustering is a method used to identify heterogeneity among groups and homogeneity within a group of patients. Without a prespecified outcome entry, the resulting model deciphers patterns that may not be disclosed using traditional methods. This is the first time such clustering analysis is applied in identifying unique subgroups at high risk for periodontitis in National Health and Nutrition Examination Surveys (NHANES 2009 to 2014 data sets using >500 variables.

**Methods:** Questionnaire, examination, and laboratory data (33 tables) for >1,000 variables were merged from 14,072 respondents who underwent clinical periodontal examination. Participants with ≥6 teeth and available data for all selected categories were included (N = 1,222). Data wrangling produced 519 variables. k-means/modes clustering (k = 2:14) was deployed. The optimal k-value was determined through the elbow method, formula = $\sum (x_i^2) - ((\sum x_i)2 /n)$. The 5-cluster model showing the highest variability (63.08%) was selected. The 2012 Centers for Disease Control and Prevention/American Academy of Periodontology (AAP) and 2018 European Federation of Periodontology/AAP periodontitis case definitions were applied.

**Results:** Cluster 1 (n = 249) showed the highest prevalence of severe periodontitis (43%); 39% self-reported "fair" general health; 55% had household income <$35,000/year; and 48% were current smokers. Cluster 2 (n = 154) had one participant with periodontitis. Cluster 3 (n = 242) represented the greatest prevalence of moderate periodontitis (53%). In Cluster 4 (n = 35) only one participant had no periodontitis. Cluster 5 (n = 542) was the systemically healthiest with 77% having no/mild periodontitis.

**Conclusion:** Clustering of NHANES demographic, systemic health, and socioeconomic data effectively identifies characteristics that are statistically significantly related to periodontitis status and hence detects subpopulations at high risk for periodontitis without costly clinical examinations.

KEYWORDS
chronic periodontitis, cluster analysis, dental health surveys, knowledge discovery, patient reported outcome measures

  

## 1 | INTRODUCTION

Chronic periodontitis, a microbe-initiated and host-mediated inflammatory disease includes periodontal attachment loss in susceptible individuals.[1] The disease is estimated to affect 42.2% of the dentate US population aged 30 to 79 years[2] with about 75% of seniors affected.[3] Severe periodontitis is the sixth most prevalent chronic disease in the world[4] affecting 11% of adults.[2,4] Periodontitis is a major cause of tooth loss and is linked with multiple health determinants, such as medical conditions (hypertension, atherosclerosis, diabetes mellitus (DM), obesity, other inflammatory diseases), physiological, dental, dietary, behavioral, socioeconomic, and environmental factors,[5–10] some of which are reciprocal.[6,11,12] Because periodontitis can be prevented, treated, or managed, it is imperative to identify high-risk population groups for prevention, improved clinical management, and administration of public health interventions.[13,14]

Decreasing the proportion of adults with moderate-to-severe periodontitis remains an objective of Healthy People 2030.[15] The National Health and Nutrition Examination Surveys (NHANES) conducted by the National Center for Oral Health Statistics (NCHS) of Centers for Disease Control and Prevention (CDC) collect data for population based-surveillance of health and disease among nationally representative population samples in the United States. For the first time, the protocols for the three 2-year cycles 2009 to 2010, 2011 to 2012, and 2013 to 2014 – hereafter collectively referred to as "NHANES 2009 to 2014"–included a full-mouth periodontal examination (FMPE) at six sites around all non-third molar teeth[2] as well as survey items for self-report of periodontitis. Both the probing depth (PD) and the distance from the periodontal margin to the cemento-enamel junction (CEJ) were recorded in millimeters. The resulting data set was selected for our study because NHANES protocols have not included clinical assessment of periodontal health since 2014.

Therefore, this publically accessible repository is valuable for exploring factors potentially intertwined with periodontitis. Data mining is applied to extract "useful" knowledge from large data sets. This process allows for deciphering meaning through data processing and analysis.[16,17]

The objective of this study was to explore the feasibility of applying data-driven approaches to identify unique subgroups with periodontitis by investigating >500 variables in the NHANES 2009 to 2014 data set related to periodontal clinical parameters and self-reported periodontitis, in the presence of health-related and socioeconomic factors. Income, education, age, sex, and poverty have previously been investigated and were found to be implicated with the prevalence of periodontitis.[30,31] Therefore, it is imperative to assess periodontitis in that context.

Cluster analysis was used to reveal class similarities, while maximizing heterogeneity among groups. In healthcare, clustering can help identify participants and groups most in need of targeted interventions.[18,19] This approach can serve to supplement "clinical judgement" by taking into perspective several variables commonly collected for recording in medical charts, but routinely limited to individual interpretation by the medical provider, which invariably could result in biased interpretation depending on provider factors. In contrast, a more automated and consistent way of handling such data could result in individualizing risk assessment and subsequent potential intervention. The "k-means" clustering is a method that uses vector quantization for grouping elements. It is an unsupervised algorithm that creates inferences from big data sets using only input variables without referring to pre-defined outcomes. The "k-modes" is an extension of k-means that instead of vectors and distances uses dissimilarities to cluster (or group) observations. There is no prior knowledge of the resulting groups; and therefore, k-means/k-modes enable grouping of observations by periodontitis categories, based on all aforementioned variables into distinct categories.

## 2 | MATERIALS AND METHODS

### 2.1 | Data collection: NHANES

NHANES was conducted in agreement with the Helsinki Declaration of 1975,[20] as most recently revised in 2013.[21] NHANES are cross-sectional in design, based on multistage stratification, and clustering of the US civilian, non-institutionalized population. The CDC's National Center for Health Statistics Ethics Review Board approved the oral health data collection protocols, and all survey participants provided written informed consent.[22] Mobile examination centers that contained space for clinical examinations, sampling of body fluids, blood pressure measurement, and interviews were used.

### 2.2 | Data extraction, inclusion, and exclusion criteria

This report follows strengthening the reporting of observational studies in epidemiology (STROBE) guidelines.[23]

The data analyzed are publically available, so no approval from any institutional review board was needed.

Each participant's barcode corresponded to a unique "Sequence ID." Participants eligible for clinical

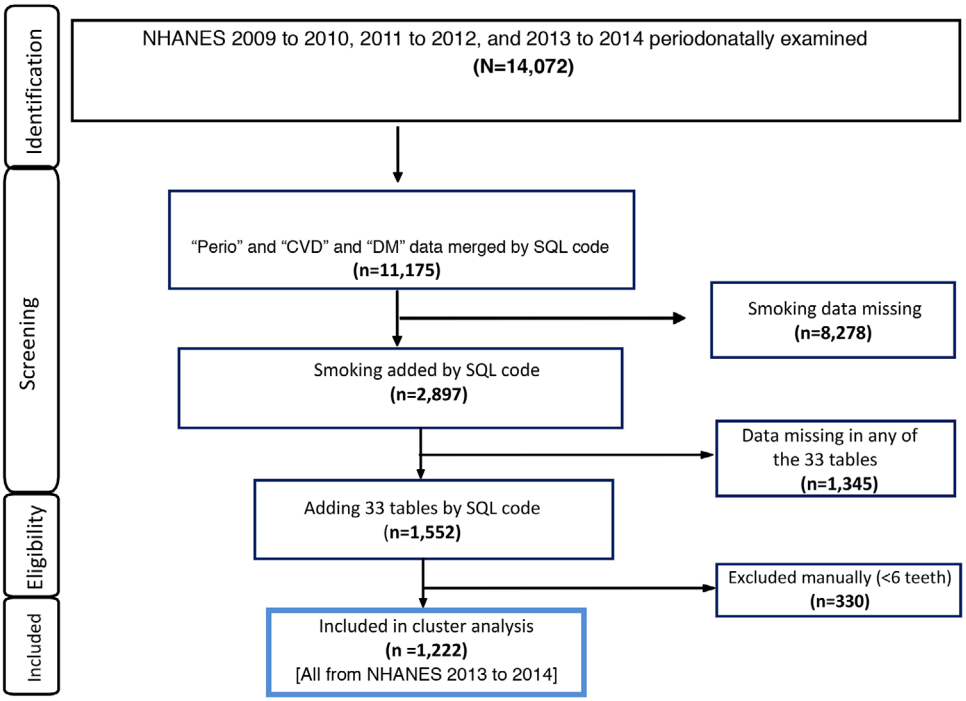


**FIGURE 1** Study participant selection from the three 2-year National Health and Nutrition Examination Survey cycles 2009 to 2010, 2011 to 2012, and 2013 to 2014 ("NHANES 2009 to 2014")

*Note.* DM, diabetes; N/n, number; SQL, structured query language (a programming language)

periodontal examination included those who 1) underwent medical history screening, 2) were aged 30 to 79 years, 3) had natural teeth, 4) were not in need of prophylactic antibiotics, and 5) provided informed consent for the oral examination. Examiners were calibrated dental hygienists (2009 to 2010) or dentists (2011 to 2014). Clinical measurements were performed as described in Methods S1A in online *Journal of Periodontology* (*JOP*). A total number of 14,072 IDs were marked as records (2009 to 2014). Thirty-three tables were selected from NHANES 2009 to 2014 l aboratory, questionnaire, and clinical data related to subjects' behavior, systemic condition(s), demographic, socioeconomic statuses, and oral health, based on previously identified periodontitis risk factors.[1,5] Variables pertaining to periodontal health status, cardiovascular status, smoking, DM, obesity, arthritis, and habits were selected. Data on demographic and socioeconomic status were collected. For systemic conditions, clinical, self-reported as well as laboratory data were used for a more comprehensive overall evaluation.

Participants whose data were excluded were those who had <6 natural teeth (to ensure adequate representation of dentition), or had data missing in any of the 33 merged data tables, including periodontal examination data. The selection process for inclusion in our study is presented in Figure 1.

## 2.3 | Application of periodontitis case definitions

Clinical attachment loss (AL) was calculated as the difference between PD and CEJ. The 2012 CDC/American Academy of Periodontology (AAP) periodontitis case definitions were applied to categorize participants with no, mild, moderate, or severe periodontitis,[24] respectively (Methods S1B in online *JOP*). Categorization into the group with severe periodontitis required ≥2 interproximal sites with clinical AL ≥6 mm (not on same tooth) and ≥1 interproximal site with PD ≥5 mm.[24]

Moreover, the European Federation of Periodontology (EFP)/AAP 2018 periodontitis classifications were applied to the best of our ability, given the availability of relevant data.[1] The disease stage was assigned to each participant based on clinical AL at 6 sites/tooth to differentiate Stage I [1 to 2 mm], II [3 to 4 mm], and III/IV [≥5 mm]; PD in Stage I [≤4mm], Stage II [≤5mm], and Stage III and IV [≥6 mm]). Participants could not be classified as either Stage III or Stage IV because the reason behind tooth loss was not included in the NHANES data set. No radiographic images, or information regarding factors pertaining to local periodontitis complexity were available in this data set. Consequently, an attempt to apply the 2018 EFP/AAP case definitions was made with major limitations, such as staging

severity being assessed based on only interproximal clinical AL.

For grading, only data on the grade modifiers DM ("glycated hemoglobin A1c" and "DM diagnosis") and smoking ("daily smokers") were available. Methods S1B and Table S1 in online *JOP* provide a detailed overview of which data necessary for correct allocation of the EFP/AAP classification are available.

## 2.4 | Data analysis

The selected data tables were merged for analysis (see Methods S1A in online *JOP*). Columns with a majority of "null"/undefined values were not included in training the model. Out of >1,000 variables, 519 were included in the cluster analysis. As mentioned, k-modes, a variation of k-means, was used for clustering, whereby participants are grouped into a specified number (k) of clusters.[25] Off-the-shelf R libraries were used for clustering. The elbow method simulated k-means clustering for values from 2 to 14 (k = 2:14). For each k, a score was computed using the total within-cluster simple matching distance, namely the sum of squared estimate of errors (SSE) via the following formula:

$$\mathrm{SSE} = \sum_{n=1}^{n} \left(x_1^2\right) - \sum\nolimits_{n=1}^{n} \left(x_i\right)^2 / n,$$

where xi is the i'th (sample) value of variable x and n the number of observations (participants), to interpret how well different k-values and their corresponding models fit the data. These results were plotted in an elbow diagram (see Figure S1 in online *JOP*). The selected number of clusters (k = 5) yielded a small SSE. SSE tends to decrease toward 0 as k increases. The value of the sum of squares error is zero (SSE = 0) when k = number of data points in the data set where each data point is its own cluster. The goal was to select a small value of k that still had a low SSE. The model learns patterns in the data set and allocates each participant to exactly one cluster. Because of the high number of variables, visual assessment of heterogeneity of clusters in 3 dimensions (x-, y-, z-axes) was a challenge (see Figure S2 in online *JOP*). Validation of the model relied on inter-class variability (63.08%) and post-scoring analysis. Based on oral variables included in clustering, principal component analysis (PCA) was conducted to reduce the dimensions, accounting only for the greatest variance/variability.

The unsupervised method of clustering was executed prior to any determination of values,' thresholds, and application of either periodontal classification. For included variables (>500) and analysis of clusters' outcome, cutoff values used for demographic and socioeconomic factors were those originally used in the NHANES data set and medical thresholds were defined based on standard national health values, such as those determined by the American Diabetes Association for (HbA1c).[26]

In an attempt to further assess the identified outcome systemic factors related to periodontal health, multivariance analysis was conducted using obesity, smoking, and DM as examples. These variables were: 1) body mass index (BMI), 2) smoking, 3) systolic and diastolic blood pressure, 4) pre-DM/DM, and 5) elevated cholesterol levels.

Different forms of bias were mitigated. Confirmation bias, for example, was dealt with by assigning equal weights to all variables entered into the model. Other forms of bias are discussed later.

### 2.4.1 | Model validity evaluation

To further evaluate the model, we conducted a complete second execution of clustering, including re-training and re-scoring. The second run used a subset of the variables: age, sex, income, and education, as well as smoker/non-smoker, number of cigarettes smoked daily, and HbA1c level, along with clinical periodontal data; that is, variables that are agreed upon in the literature as correlated to periodontitis. As an outcome, Cluster 1 was distinct from other clusters, as it had no severe, moderate, or mild periodontitis. However, other clusters from the same perspective were not clearly distinct (i.e., their inter-cluster values were high resulting in an overlap between the clusters; or clusters included a small number of participants). Such outcomes mean that variables passed to the model are not sufficient for it to identify unique patterns in all clusters. This was not a surprise because it is known that unsupervised models require big data, namely more than included in this second re-run. Accordingly, the first execution was deemed more successful because it is more reasonable and includes all variables.

## 3 | RESULTS

From the initially identified 14,072 participants with periodontal data, a total of 1,222 were included in our analyses (Fig. 1). All had participated in the 2013 to 2014 cycle.

## 3.1 | Allocation to the five clusters

The 1,222 included participants were allocated by the model to exactly 1 of 5 clusters, resulting in the following

cluster sizes: Cluster 1 (n = 249), Cluster 2 (n = 154), Cluster 3 (n = 242), Cluster 4 (n = 35), and Cluster 5 (n = 542).

### 3.2 | Categories of periodontitis in the five clusters

The model resulted in grouping participants with the greatest proportion of no periodontitis in Cluster 2, mild periodontitis in Cluster 5, moderate periodontitis in Cluster 3 followed by Cluster 4, and severe periodontitis in Cluster 1 (Table 1). The proportion with total or any (mild, moderate, or severe) periodontitis in Cluster1 was 65.0%, namely 6.0% mild, 16 .1% moderate, and 43% severe periodontitis, leaving about one-third (34.9%) with no periodontitis. The corresponding results in Cluster 5 were 78.7% with total periodontitis (55.5% mild, 20.9% moderate, and 2.3% severe), leaving about one-fifth (21.3%) with no periodontitis. In reference to all other included variables, Figures 2A and 2B and Figure S3A through S3C in online *JOP* display the top influential variables, arranged in order of magnitude per cluster.

### 3.3 | Demographic and socioeconomic factors

The non-Hispanic White race/ethnic group prevailed in all clusters, followed by non-Hispanic Black group in Clusters 1, 2, and 3. The sex distribution was about even, except in Cluster 4 in which males constituted >70% (71.42%) (see Table S2 in online *JOP*). In Clusters 1 through 4, >40% reported annual family income of <$35,000 US dollars (USD). Only in Cluster 5, the majority of participants earned between $35,000 USD and $99,999 USD. Two-thirds of Cluster 1 had attained high school or less education versus half in Cluster 5. The most prevalent age group in all clusters was 50 to 69 years, most explicitly in Cluster 1 (71.08%), where age 50 to 69 years was the eighth most influential variable (Fig. 2A).

### 3.4 | Periodontal status, oral health, and recommendations of care

Cluster 1 followed by Cluster 4 contained the greatest proportions of participants with severe periodontitis; shown in Table 2. In accordance with EFP/AAP case definitions, this group also showed the greatest prevalence of Stage IV periodontitis (52%) (Fig. 3). In Cluster 1, 109 of identified smokers had periodontitis and were classified within good, fair, and poor general self-reported general health. A large subset of the same group (80 participants) had several decayed teeth. This group also had the greatest proportion of grade C risk modifiers (17%), based on the number of cigarettes smoked daily and the HbA1c level. One participant in Cluster 2 had periodontitis (45-year-old female, multiple sites with clinical AL ≥4 mm, anti-hypertensive medication, and smoked >10 cigarettes a day), and 93% had Stage I periodontitis.

More than half (53%) of Cluster 3 had moderate periodontitis and 75% responded "Yes" to the question, "Do you think you might have gum disease?" In Clusters 3 and 4, 98% had some form of periodontitis, that is total periodontitis. Additionally, Cluster 4 had the greatest proportion of Stage III periodontitis (83%) followed by Cluster 3 (58%). Finally, 77% of Cluster 5 had only no or mild periodontitis. Three-quarters (76%) of these subjects had Stage II periodontitis.

Risk modifiers were included in grading. Participants were categorized by either DM together with number of cigarettes or DM exclusively, depending on data availability. Only 2.6% of all participants could not be graded by either factor. Of 985 subjects considered as Grade A, 461 subjects were determined by both modifiers and 524 only by diabetes variables.

Furthermore, a total of 85% of participants in Cluster 4 were thought to have oral hygiene issues, followed by 65% in Cluster 3 and 45% in Cluster 1. Interestingly, while >40% in Clusters 3 and 4 had decayed teeth, 82% in Cluster 5 had none. Additionally, Cluster 4 contained the greatest proportion (46%) advised to see a dentist either immediately due to acute injuries or within 2 weeks of examination due to "chronic pain, gum issues, or for counseling."

The scatterplot in Figure 4 represents the 5-cluster model derived by the PCA based on dental and care recommendation variables. It shows homogeneity and heterogeneity among individuals in the same cluster and between clusters.

### 3.5 | Chronic conditions

DM, coronary heart disease, rheumatoid arthritis (RA), and hypercholesterolemia were among the most alternating conditions among clusters. Of all included participants, 13% had DM (HbA1c ≥6.5%). Among those with DM, elevated fasting blood glucose level (>100 mg/dL) was found in >65% of Clusters 1 and 4. Nearly 1 in 10 (8.6%) of those suffering from DM had HbA1c levels of >7%, that are uncontrolled DM. The greatest proportions of them (15%) were found in Clusters 1 and 4. Moreover, 40% of Cluster 1 had pre-DM defined as HbA1c between 5.7% and 6.4% (Fig. 2A; Table S3 in online *JOP*). Interestingly, participants in Clusters 2 and 3 showed similar proportions of pre-DM according to HbA1c but not to the 2-hour glucose

JOURNAL OF Periodontology



**TABLE 1** Proportions of participants within the top influential variables sorted from the highest (green) to the lowest (red) value (Heatmap) displayed in descending order according to Cluster 1–National Health and Nutrition Examination Survey (NHANES) 2013 to 2014 (N = 1,222).*

| Diabetes related variables | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| Do not take insulin | 91.2% | 94.2% | 94.6% | 97.1% | 100.0% |
| Fasting blood glucose ≥100 mg/dL | 65.5% | 54.5% | 63.2% | 65.7% | 49.6% |
| Answered "No" to DM | 57.0% | 50.0% | 55.8% | 60.0% | 53.5% |
| OGTT <140 | 52.2% | 44.8% | 52.5% | 51.4% | 64.9% |
| Close relatives with DM | 43.0% | 50.0% | 44.2% | 40.0% | 46.7% |
| HbA1c - 5.7-6.4% | 41.4% | 33.8% | 37.6% | 20.0% | 27.7% |
| HbA1c - ≤ 5.6% | 39.4% | 50.0% | 43.0% | 62.8% | 64.5% |
| OGTT >200 mg/dL | 32.5% | 3.3% | 5.4% | 2.8% | 4.6% |
| Officially diagnosed with DM | 20.9% | 20.1% | 19.4% | 17.1% | 10.3% |
| On diabetic pills | 14.1% | 14.9% | 15.7% | 14.3% | 8.1% |
| OGTT - 140-200 mg/dL | 11.6% | 14.9% | 18.6% | 20.0% | 19.2% |
| HbA1c - >8% | 10.0% | 7.1% | 5.8% | 8.6% | 2.6% |
| HbA1c - 6.5-7% | 4.0% | 7.8% | 8.2% | 2.9% | 3.5% |
| On insulin | 8.8% | 5.8% | 5.4% | 2.9% | 0.0% |
| Reported complications (exl. retinopathy) | 6.0% | 3.9% | 2.9% | 5.7% | 2.0% |
| HbA1c -7.1-8% | 5.2% | 1.30% | 5.40% | 5.70% | 1.70% |

| Demographics | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| High school graduates or less | 71.8% | 48.7% | 53.7% | 54.2% | 26.6% |
| Age 50-69 years | 71.1% | 68.8% | 65.3% | 48.6% | 53.5% |
| Annual family income $0-$34,999 | 55.4% | 46.8% | 45.9% | 43.7% | 22.9% |
| Male Sex | 52.2% | 44.8% | 49.6% | 71.4% | 44.8% |
| Non-Hispanic white | 35.3% | 39.0% | 34.3% | 14.3% | 48.0% |
| Annual income $35,000-$99,999 | 32.5% | 37.1% | 39.7% | 36.3% | 41.7% |
| Female Sex | 27.8% | 55.1% | 50.4% | 28.5% | 55.2% |
| Non-Hispanic Black | 27.3% | 26.0% | 28.1% | 14.3% | 11.8% |
| Age 30-49 | 26.9% | 22.7% | 28.5% | 42.9% | 42.1% |
| Some college degree | 20.1% | 30.5% | 28.9% | 34.3% | 31.4% |
| Mexican American | 13.6% | 12.3% | 15.7% | 25.7% | 11.3% |
| Other Hispanic | 13.6% | 8.4% | 8.7% | 14.3% | 9.4% |
| Other-multiracial | 10.2% | 14.3% | 13.2% | 31.4% | 19.6% |
| College or above | 8.0% | 20.8% | 17.4% | 11.5% | 42.1% |
| Annual family income ≥$100,000 | 6.0% | 11.6% | 11.6% | 11.4% | 32.8% |
| Age 70 -79 years | 2.0% | 4.6% | 5.4% | 5.7% | 4.1% |

| Periodontitis | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| Total Periodontitis | 65.0% | 0.7% | 98.8% | 97.1% | 78.7% |
| Severe Periodontitis | 43.0% | 0.7% | 24.8% | 31.4% | 2.3% |
| No Periodontitis | 34.9% | 99.4% | 1.2% | 2.9% | 21.3% |
| Moderate Periodontitis | 16.1% | 0.0% | 52.9% | 51.4% | 20.9% |
| Mild Periodontitis | 6.0% | 0.0% | 21.1% | 14.3% | 55.5% |

(Continues)

**TABLE 1** (Continued)

| Oral health and recommendation of care | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| Gum disease | 62.3% | 8.4% | 78.5% | 40.0% | 19.7% |
| Recommendation of care - other | 62.3% | 65.6% | 89.7% | 54.3% | 98.2% |
| Decayed teeth | 43.8% | 20.1% | 47.5% | 45.7% | 16.9% |
| Oral hygiene - Yes | 53.4% | 9.0% | 64.5% | 85.7% | 21.4% |
| Do not floss | 49.8% | 53.0% | 24.8% | 57.2% | 14.2% |
| Partial dentures or plates | 30.1% | 7.8% | 9.9% | 8.6% | 0.4% |
| Flossing -(7days a week) | 25.7% | 22.1% | 46.3% | 25.7% | 41.3% |
| Flossing -(1-4 days a week) | 24.5% | 24.9% | 28.9% | 17.1% | 44.5% |
| Had treatment for gum disease | 22.1% | 22.7% | 27.3% | 22.9% | 23.8% |
| Recommendation of care-immediate/urgent | 21.3% | 3.9% | 10.3% | 45.7% | 1.8% |
| Told to have bone loss | 18.8% | 16.3% | 20.3% | 22.9% | 9.2% |
| **Habits** | | | | | |
| **1) Smoking** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** |
| At least 100 cigarettes a life | 77.7% | 55.8% | 51.7% | 45.7% | 24.4% |
| Smokers at home - none | 61.8% | 66.8% | 65.7% | 99.0% | 85.4% |
| Cotinine >102 mg/dL | 48.2% | 33.0% | 43.4% | 31.4% | 12.5% |
| Cotinine <1 mg/dL | 42.1% | 58.0% | 51.8% | 60.0% | 82.8% |
| Smokers at home | 38.2% | 33.2% | 34.3% | 1.0% | 14.6% |
| Daily smokers | 37.8% | 25.3% | 23.6% | 28.6% | 24.4% |
| Cotinine 10-101 mg/dL | 7.4% | 7.0% | 2.8% | 5.7% | 3.3% |
| Cotinine 1-10 mg/dL | 3.6% | 2.0% | 2.0% | 2.9% | 1.4% |
| **2) Lifestyle** | **Cluster1** | **Cluster2** | **Cluster3** | **Cluster4** | **Cluster5** |
| Trying to control weight | 63.5% | 59.7% | 70.2% | 62.9% | 66.8% |
| Attempting to reduce fat | 61.0% | 53.2% | 60.7% | 60.0% | 61.8% |
| Attempting to reduce salt | 59.8% | 54.5% | 57.8% | 60.0% | 66.7% |
| Attempting to increase exercise | 57.8% | 55.2% | 64.9% | 48.6% | 9.8% |
| **3) Other** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** |
| Daily marijuana smokers | 26.9% | 13.0% | 22.3% | 17.1% | 11.6% |
| Ever used meth/cocaine/heroin | 21.3% | 18.2% | 18.6% | 11.4% | 11.8% |
| Answered "No" to meth/cocaine | 15.2% | 16.2% | 4.4% | 14.2% | 6.5% |

HbA1c, glycated hemoglobin; OGTT, 2-hour glucose tolerance test; SPF, socio-position factors.

*The table is arranged in ascending order of Cluster 1 and the total percentages per category add up to a 100%. For instance, the proportion with total periodontitis in Cluster 1 is 65%, categorized as 43% severe, 16% moderate, 6% mild, and 34.9% no periodontitis (totaling 100% of 65%).

test or fasting blood glucose level (see Table S3 in online *JOP*).

Regarding participants' cardiovascular health status, the majority in Cluster 1 took anti-hypertensive medications and >60% had been told they had elevated blood pressure. Having been told of having coronary heart disease represented about 10% in all groups and did not play a significant role in characterizing clusters.

Most participants who had smoked ≥100 cigarettes in their lifetime were grouped into Cluster 1. "Active smoker" shown by >10 mg/dL cotinine in urine samples accounted for >50% in Clusters 1 and 3. Current daily smokers constituted 38% of Cluster 1, the greatest prevalence among all clusters. The variable "smoked ≥100 cigarettes in your lifetime" was largely representing past smokers in Cluster 2, shown by cotinine levels <1 mg/dL, reflecting the lack of active and environmental smoke exposure.

More than 80% in Cluster 5 fell within normal range of laboratory testing for cotinine, and 85% stated there were no smokers in the household. Low-density lipoprotein (LDL) levels were highest in Cluster 4. Clusters 1 and 4 featured the greatest proportions of individuals

**TABLE 2** Self-reported periodontal disease, clinically determined oral health, and recommendations of dental care by clusters–NHANES 2013 to 2014 (N = 1,222)

| Characteristic | Cluster 1 (n, %) | Cluster 2 (n, %) | Cluster 3 (n, %) | Cluster 4 (n, %) | Cluster 5 (n, %) | Total n (%) |
|---|---|---|---|---|---|---|
| Total/All | 249 (20.4%) | 154 (12.6) | 242 (19.8) | 35 (2.9%) | 542 (44.3%) | 1,222 (100%) |
| No. of teeth present mean (±SE) | 13.0 (±10.4) | 19.0 (±10.2) | 23.5 (±10.5) | 26.0 (±10.3) | 25.0 (±10.2) | 20.6 (±10.3) |
| Think you have gum disease? | | | | | | |
| Yes | 101 (40.6%) | 9 (5.9%) | 167 (69.0%) | 14 (40.0%) | 107 (19.7%) | 398 (32.6%) |
| No | 0 (0%) | 0 (0%) | 0 (0%) | 1 (2.9%) | 0 (0%) | 1 (0.1%) |
| NR | 148 (59.4%) | 145 (94.1%) | 75 (31.0%) | 20 (57.1%) | 435 (80.3%) | 823 (67.3%) |
| Ever been told you have lost bone around your teeth? | | | | | | |
| Yes | 47 (18.8%) | 25 (16.3%) | 49 (20.3%) | 8 (22.9%) | 50 (9.2%) | 179 (14.6%) |
| No | 201 (80.8%) | 128 (83.1%) | 190 (78.5%) | 27 (77.1%) | 488 (90.0%) | 1,034 (84.6%) |
| N/R | 1 (0.4%) | 1 (0.6%) | 3 (1.2%) | 0 (0%) | 4 (0.8%) | 9 (0.8%) |
| Ever had treatment for gum disease? | | | | | | |
| Yes | 55 (22.1%) | 35 (22.7%) | 66 (27.3%) | 8 (22.9%) | 129 (23.8%) | 293 (24.0%) |
| No | 191 (76.7%) | 117 (76.0%) | 176 (72.7%) | 27 (77.1%) | 413 (76.2%) | 924 (75.6%) |
| N/R | 3 (1.2%) | 2 (1.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 5 (0.4%) |
| Oral hygiene* | | | | | | |
| Yes | 133 (53.4%) | 14 (9.0%) | 156 (64.5%) | 30 (85.7%) | 116 (21.4%) | 449 (36.7%) |
| No | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| N/R | 116 (46.6%) | 140 (91.0%) | 86 (35.5%) | 5 (14.3%) | 426 (78.6%) | 773 (63.3%) |
| Flossing/using other devices | | | | | | |
| 7 days/week | 61 (24.5%) | 38 (24.9%) | 70 (28.9%) | 6 (17.1%). | 241 (44.5%) | 416 (34.0%) |
| 1 to 4 days | 64 (25.7%) | 34 (22.1%) | 112 (46.3%) | 9 (25.7%) | 224 (41.3%) | 443 (36.3%) |
| N/R | 124 (49.8%) | 82 (53.0%) | 60 (24.8%) | 20 (57.2%) | 77 (14.2%) | 363 (29.7%) |

(Continues)

**TABLE 2** (Continued)

| Characteristic | Cluster 1 (n, %) | Cluster 2 (n, %) | Cluster 3 (n, %) | Cluster 4 (n, %) | Cluster 5 (n, %) | Total n (%) |
|---|---|---|---|---|---|---|
| Decayed teeth | | | | | | |
| Yes | 109 (43.8%) | 31 (20.1%) | 115 (47.5%) | 16 (45.7%) | 92 (16.9%) | 363 (29.7%) |
| No | 140 (56.2%) | 123 (79.9%) | 127 (52.5%) | 19 (54.3%) | 448 (82.7%) | 857 (70.1%) |
| N/R | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (0.4%) | 2 (0.2%) |
| Periodontitis category (CDC/AAP case definitions[24])† | | | | | | |
| No | 87 (35.0%) | 153 (99.3%) | 3 (1.2%) | 1 (2.9%) | 116 (21.4%) | 360 (30.0%) |
| Mild | 15 (6.0%) | 0 (0%) | 51 (21.1%) | 5 (14.3%) | 301 (55.6%) | 372 (30.4%) |
| Moderate | 40 (16.0%) | 0 (0%) | 128 (52.9%) | 18 (51.4%) | 113 (20.8%) | 299 (24.4%) |
| Severe | 107 (43.0%) | 1 (0.7%) | 60 (24.8%) | 11 (31.4%) | 12 (2.2%) | 191 (15.2%) |
| Total | 162 (65.0%) | 1 (0.7%) | 239 (98.8%) | 34 (97.1%) | 426 (78.7%) | 862 (70.0%) |
| Periodontitis category (EFP/AAP case definitions[1])† | | | | | | |
| Stage I | 9 (3.6%) | 145 (93.2%) | 0 (0%) | 0 (0%) | 61 (11.3%) | 213 (17.4%) |
| Stage II | 81 (32.5%) | 10 (6.1%) | 73 (30.2%) | 6 (17.1%) | 409 (75.5%) | 579 (47.4%) |
| Stages III/IV | 159(63.8%) | 1 (0.7%) | 169 (69.8%) | 29 (82.9%) | 72 (13.3%) | 430 (35.2%) |
| Grade A‡ | 180 (72.3%) | 123 (79.9%) | 186 (76.9%) | 28 (80.0%) | 468 (86.3%) | 985 (80.6%) |
| Grade B | 22 (8.8%) | 13 (8.4%) | 16 (6.6%) | 2 (5.7%) | 30 (5.5%) | 83 (6.8%) |
| Grade C | 43 (17.3%) | 11 (7.2%) | 32 (13.2%) | 5 (14.3%) | 31 (5.7%) | 122 (10.0%) |
| Not determined | 4 (1.6%) | 7 (4.5%) | 8 (3.3%) | 0 (0%) | 13 (2.5%) | 32 (2.6%) |
| Recommendation for professional dental care | | | | | | |
| See a dentist | 53 (21.3%) | 6 (3.9%) | 25 (10.3%) | 16 (45.7%) | 10 (1.8%) | 110 (9.0%) |
| Other§ | 156 (62.7%) | 101 (65.6%) | 217 (89.7%) | 19 (54.3%) | 532 (98.2%) | 1,025 (83.9%) |
| N/R | 40 (16.0%) | 47 (30.5%) | 0 (0%) | 0 (0%) | 0 (0%) | 87 (7.1%) |

AAP, American Academy of Periodontology; CDC, Centers for Disease Control and Prevention; EFP, European Federation of Periodontology; N/n, number; N/R, Not reported; See a dentist, See a dentist immediately or within the next 2 weeks; total periodontitis, mild or moderate, or severe (taken together).

The subgroups of characteristics are categorized and displayed as described by CDC website in NHANES questionnaires and clinical examination.

*Oral Hygiene refers to the examiner's assessment of "other" conditions designated in referral letters. Oral hygiene, gum disease/problems are examples.

†Clinical periodontitis status defined by CDC/AAP periodontitis case definitions (Eke et al., 2012) and 2018 EFP/AAP classification. In grading, the category under "not determined" refers to those considered with borderline diabetes and who have not received a diagnosis of diabetes but HbA1c percentage is between 6.5% and 6.9%.

‡Grading was assessed based on risk modifiers, namely, smoking and diabetes. For Grade A, 524 subjects were assessed by diabetes. For Grade B, 43 subjects were based on diabetes only. Those who were not categorized could not be allocated to a group due to lack of smoking and diabetes data due to missing values in the original NHANES data set.

§Other recommendations of care included "See a dentist at your earliest convenience," and "Continue your regular routine care."

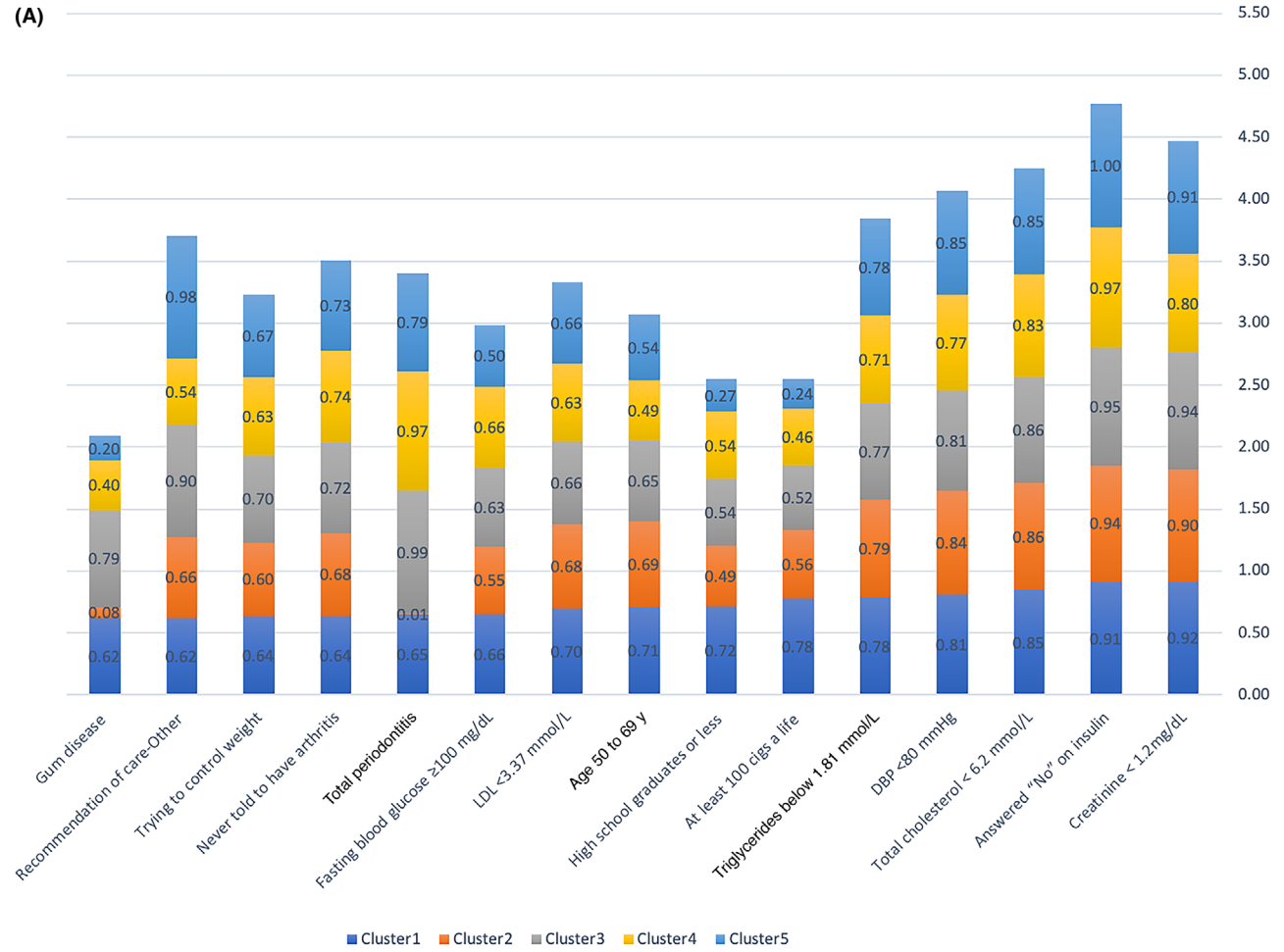


**FIGURE 2** Top variables describing **A**) Cluster 1 (unhealthier) and **B**) Cluster 5 (healthier), variables arranged in descending order of the greatest/right magnitude/influence to the lowest/left on grouping subjects within Cluster 1 (dark blue) and Cluster 5 (light blue), respectively. The 2012 CDC/AAP periodontitis case definitions are applied.[24] Please note, the corresponding graphs for Clusters 2, 3, and 4 are displayed in supplementaryFigure S3. DBP, diastolic blood pressure

with elevated (>2.2 mmol/L) triglyceride levels, contributing to their poorer overall health status. Finally, one-third of Cluster 1 members had been told they had RA.

### 3.6 | Other health conditions and habits

In Cluster 1, 39% assessed their general health as "fair," 23% had DM, almost half (48%) had elevated cotinine levels, and 33% had elevated LDL levels. Clusters 3 and 4 showed the greatest proportions rating their overall health condition "fair." Cluster 5 included the healthiest members with 30% regarding their health "very good" (Fig. 2B, see Table S3 in online *JOP*).

Obesity (BMI ≥30 kg/m$^2$) was significant to grouping participants and obese individuals constituted 35% of all clusters. Additionally, Cluster 4 had the greatest proportion of overweight people (25 ≤ BMI < 30 kg/m$^2$) with more than half (54%) affected. Multivariate analysis showed the greatest F-values for BMI and cotinine, meaning their values were highly variable and contributed strongly to the process of clustering (see Table S4 in online *JOP*).

## 4 | DISCUSSION

Considering that our five clusters only include data from 1,222 of the participants in the 2013 to 2014 NHANES cycle, our results regarding the prevalence of periodontitis defined by the CDC/AAP classifications,[2] obesity,[27] and DM are in line with the respective CDC findings.[28] Results from glucose tolerance tests and the HbA1c levels enabled detection of undiagnosed pre-DM as well as DM. This finding could support advocating for action on identifying pre-DM and early detection of disease in the dental care setting, especially for people who do not undergo routine medical health checkup,[29,30] but do visit a dental office.

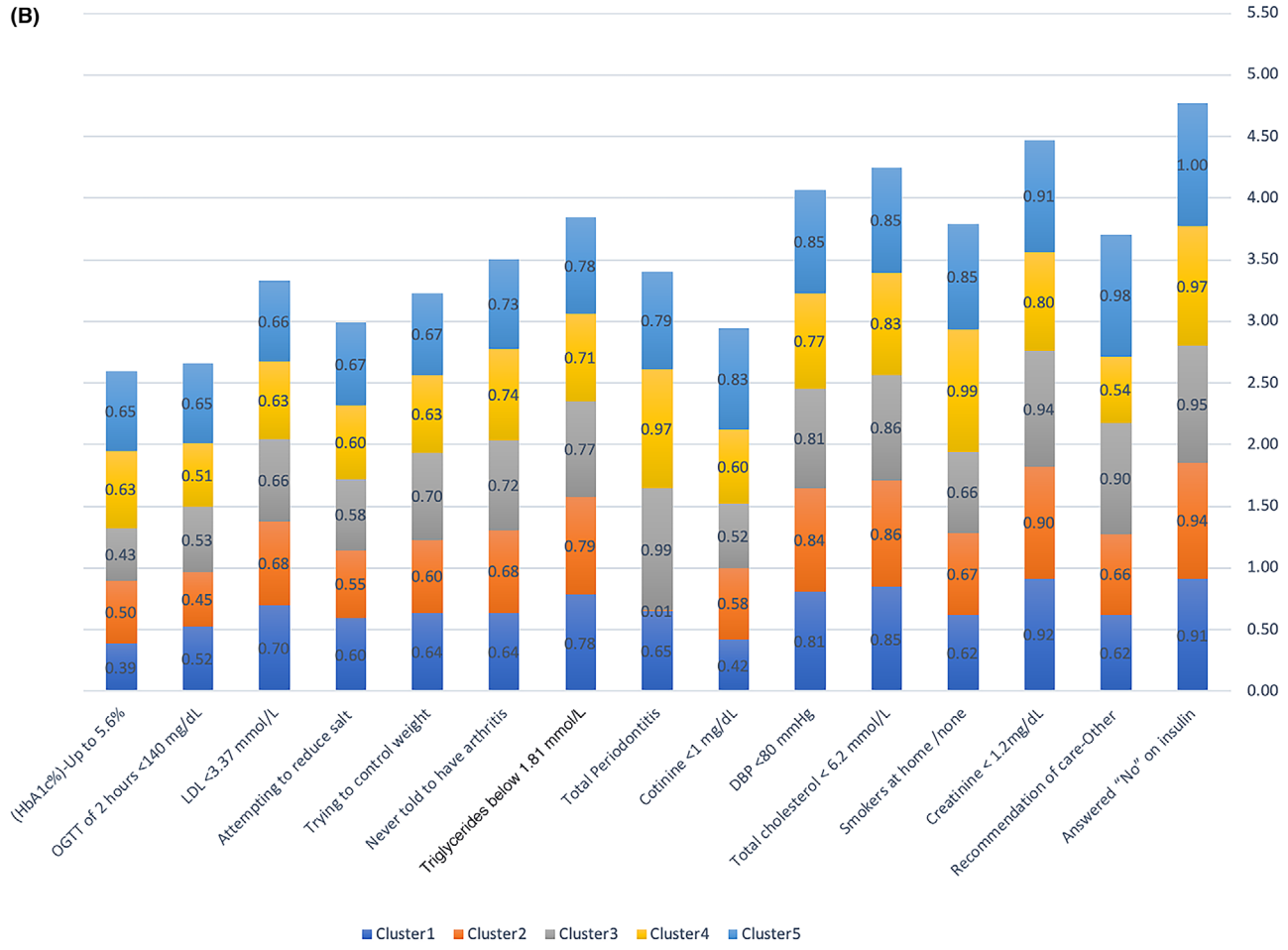


**FIGURE 2** Continued

Likewise, we confirm previous findings that show an inverse relationship between prevalence of periodontitis and socioeconomic status assessed by education attainment and family income.[31–34] This is evident in Cluster 1, in which the majority of individuals had family income >$35,000 USD and high school level of education versus Cluster 5 where 40% had college education and one-third had annual family income >$100,000 USD.

We applied the CDC/AAP periodontitis case definitions because they were designed specifically for population surveillance.[24] When attempting to translate these case definitions[24] to the 2018 EFP/AAP classification,[1] moderate and severe periodontitis defined by the former are most likely of Stage III or IV. Our data show an overall prevalence of severe periodontitis of 15%, most prominent in Cluster 1 that contains the greatest proportion of smokers. Since the majority were current smokers, had severe periodontitis, and had HbA1c levels ≥7%, they would potentially exhibit rapid progression and hence be classified as Grade C.[1] Age groups >60 years (n = 329) showed a prevalence of 22.4% versus only 0.1% severe periodontitis cases among 30- to 50-year-olds (n = 410). Previously, adults aged ≥65 years were estimated to have a 7-fold increased risk of periodontitis compared with younger groups.[3,37,38] While Cluster 4 showed the second largest group of severe periodontitis; it is likely due to oral hygiene issues as reported by dental examiners.

Despite some overlap in demographic, socioeconomic, and systemic conditions (e.g., obesity, pre-DM), Cluster 3 showed the greatest proportion with moderate periodontitis, while Cluster 2 had almost none with any periodontitis. This difference is likely due to oral hygiene issues, prevalent DM, and current smoking in Cluster 3.

## 4.1 | Strengths

Because this study is the first of its kind, there are no prior studies to which we can compare our results. Unsupervised learning models are increasingly popular in precision medicine. For example, this approach can help identify homogeneous groups and result in deciphering stronger associations between periodontitis and underlying risk factors.[39] We applied k-means clustering as such a model

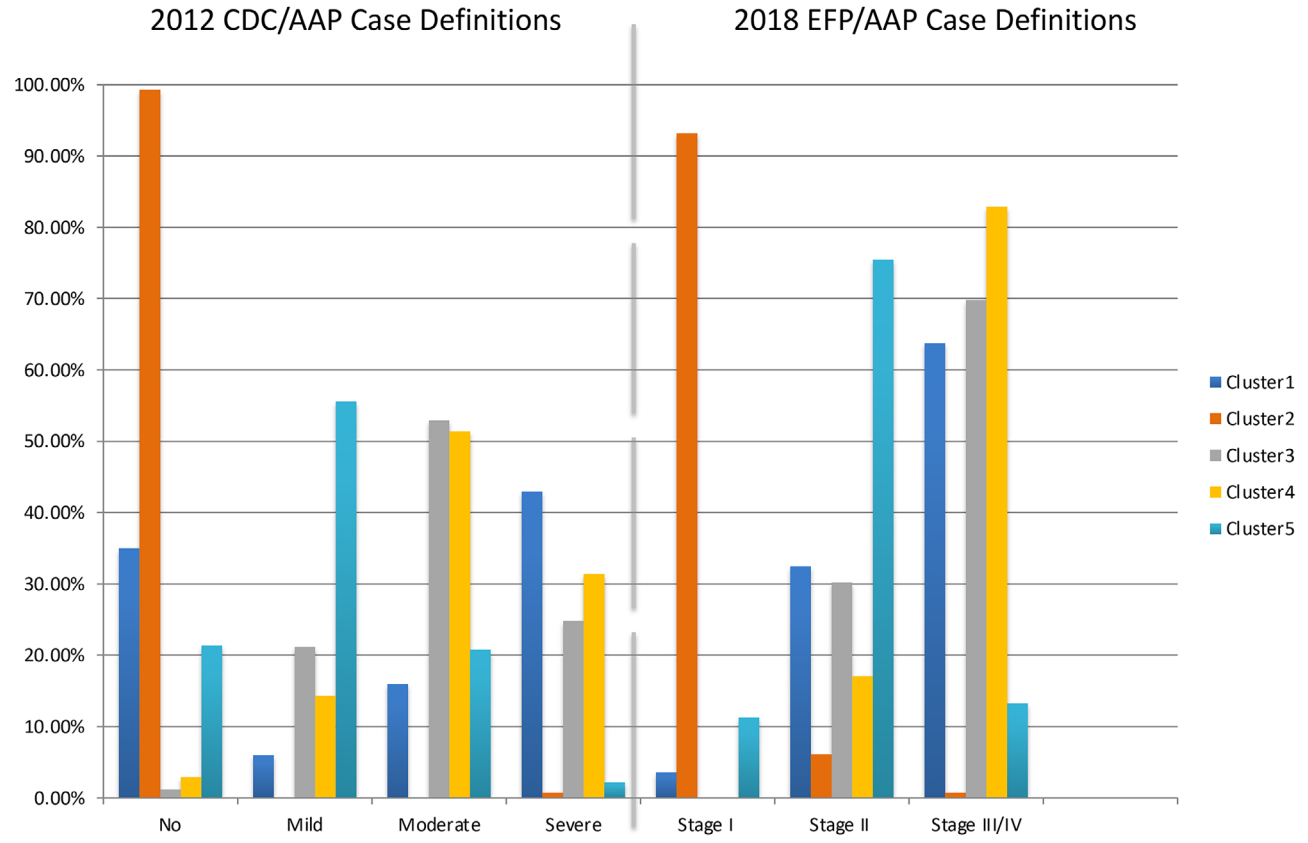


**FIGURE 3** Distribution by cluster of periodontitis prevalence according to the 2012 CDC/AAP case definitions for no, mild, moderate, and severe periodontitis[24] and Stages I, II, and III/IV according to the 2018 EFP/AAP periodontitis classification.[1] DBP, diastolic blood pressure

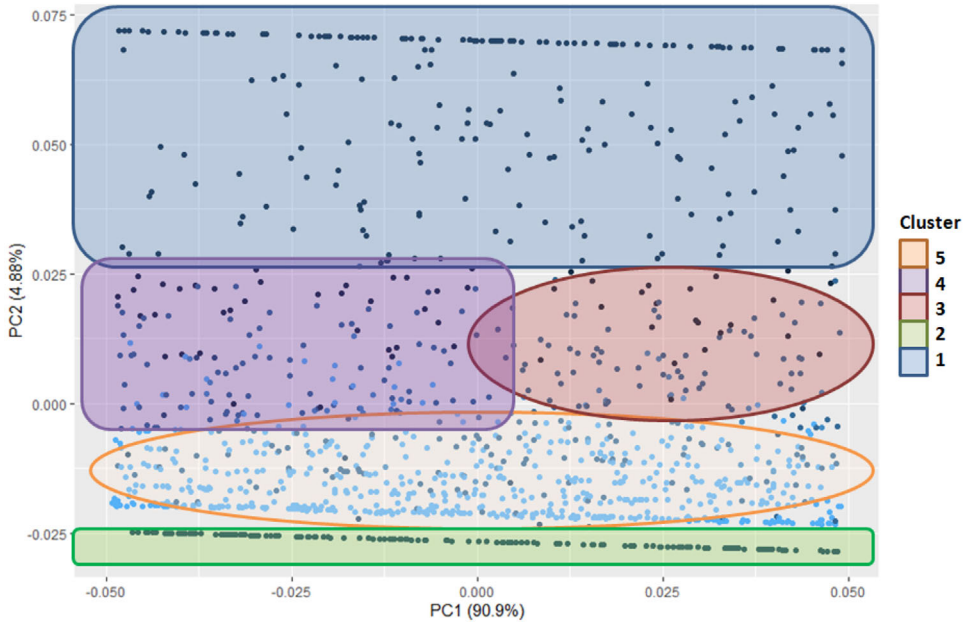


**FIGURE 4** Principal component analysis scatterplot representation of the 5-cluster model based on periodontal health variables. The 2012 CDC/AAP periodontitis case definitions are applied.[24] An R library (ggfortify) was used to consume the variables from the R engine and illustrate the clusters in a colored scheme. For example, the figure shows that Cluster 1 (shaded blue) has the highest distribution among the components space. Participants in Cluster 2 (green) are those most strongly associated with each other. Clusters 1 and 2 exhibit the furthest association in periodontal terms (most different). Clusters 2 and 5 (orange) are the periodontally closest (most similar) with >99% having no periodontitis in Cluster 2 and 77% having no or mild periodontitis in Cluster 5. Clusters 3 and 4 are intermediary to Cluster 1 and 5

that groups participants into distinct categories based on periodontal variables with no pre-determined outcomes assigned. We showed that cluster models indeed can help identify population groups at high risk for periodontitis and hope that this common disease thereby can be better prevented, identified, treated, or managed.

To the best of our knowledge, this is the first comprehensive model built via exploring a large, nationally representative database with >1,000 original variables to group periodontally examined participants.

Our comprehensive data set was extracted from the NHANES data collected on many periodontitis-related factors that were identified a priori according to their known influence on pathogenesis and progression of periodontitis.[6] The NHANES 2009 to 2014 data set is the world's largest reference to periodontitis prevalence. This is due to the extensive number of participants aged 30 to 79 years and to the application of the gold standard periodontal examination with probing at six sites around all non-third molar teeth for both PD and CEJ, ensuring that the prevalence of periodontitis can be estimated by applying various periodontitis case definitions that include clinical AL and PD, such as mean clinical AL and PD as well as proportions with PD and clinical AL above various thresholds.[2] These NHANES 2009 to 2014 data are the first to include measures from full-mouth periodontal clinical examinations, which should greatly improve the validity and reliability of estimates compared with data from earlier NHANES partial-mouth periodontal examination protocols that have underestimated the prevalence of periodontitis by up to 54%.[40]

Unlike some previous models, dental factors, that is, caries, oral hygiene, and number of teeth, were included. The threshold of having a minimum of six natural teeth was selected to ensure inclusion of representative dentate participants due to having fewer teeth might indicate hopeless periodontitis or other severe situations.

## 4.2 | Limitations

The cross-sectional study design of NHANES encompasses inherent limitations that prevent any conclusion regarding potential causality due to the simultaneous occurrence of the observed information.[41] Only the 2013 to 2014 NHANES 2-year cycle was included after merging data tables. While PD and CEJ were recorded to assess the periodontal status, there was no assessment of the presence of dental plaque nor of bleeding on probing, which potentially could limit the identification of other disease phenotypes.[39] Additionally, application of the EFP/AAP classification was limited due to availability of relevant data in the NHANES data set. Importantly, participants could not be classified distinctly into either Stage III or Stage IV, so these were merged into one category. Information on local complexity factors and radiographs are missing and thus, preventing accurate staging and grading. Only clinical AL could be regarded informative pertaining to severity. The CDC/AAP case definitions[24] define attachment loss of ≥3 mm at ≥2 interproximal sites (not on the same tooth) as mild periodontitis. In the EFP/AAP classification, Stage II indicates attachment loss of 3 to 4 mm. Overlap could exist between Stage II (EFP/AAP) and mild periodontitis (CDC/AAP). Nonetheless, the aim of using both case definitions was to allow interpretation of the resulting clusters output in a familiar context in the field. However, comparing periodontitis prevalence determined by the two sets of case definitions might not adequately represent the results of clusters at this point. The use of CDC/AAP case definitions seem more justifiable to use with NHANES data set as detailed in Supplementary Methods S1B in online *JOP*.

Furthermore, data sets from large population studies will not be complete, as some data inherently will be missing, due to participants' ineligibility or refusal to undergo certain exams, respond to certain questionnaire items, or due to the inclusion/exclusion of certain measures from study protocols that are time-sensitive or vary by survey cycle. Self-reported data were not validated. Notably, while results from such large population studies may apply in the aggregate to groups, they may not be informative at an individual level.[39]

# 5 | CONCLUSIONS

Clustering of NHANES data by integrating systemic health, demographic, and socioeconomic characteristics can effectively identify population group characteristics that are statistically significantly associated to periodontitis. Identification of such clusters can be enhanced by self-reported periodontitis measures in lieu of the extremely resource demanding clinical periodontal examination. Economical, non-intrusive clustering constitutes a low-cost alternative to identifying population groups at high risk for periodontitis, who could be targeted for preventive and therapeutic dental public health intervention.

## ACKNOWLEDGMENTS

No funding was received for this study. The authors report no conflicts of interest related to this study.

## AUTHOR CONTRIBUTIONS

Iya Ghassib: contributed to conception; study design; data extraction, preparation, analysis and interpretation; drafted and critically revised the manuscript. Feras A. Batarseh: contributed to conception; study design, data extraction, preparation, analysis, and interpretation; and critically reviewed the manuscript. Hom-Lay Wang: contributed to conception, drafted and critically reviewed the manuscript. Wenche S. Borgnakke: contributed to design, contributed to data interpretation; drafted and critically revised the manuscript. All authors gave their final approval and agreed to be accountable for all aspects of the work.

## ORCID

*Iya H. Ghassib* https://orcid.org/0000-0002-0067-377X
*Feras A. Batarseh* https://orcid.org/0000-0002-6062-2747
*Hom-Lay Wang* https://orcid.org/0000-0003-4238-1799
*Wenche S. Borgnakke* https://orcid.org/0000-0003-3593-093X

## REFERENCES

1. Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: framework and proposal of a new classification and case definition. *J Clin Periodontol*. 2018;45(Suppl 20):S149-S161.
2. 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;149:576-588.588.e1-588.e6.
3. Eke PI, Wei L, Borgnakke WS, et al. Periodontitis prevalence in adults ≥ 65 years of age, in the USA. *Periodontol 2000*. 2016;72:76-95.
4. Kassebaum NJ, Bernabe E, Dahiya M, Bhandari B, Murray CJ, Marcenes W. Global burden of severe periodontitis in 1990-2010: a systematic review and meta-regression. *J Dent Res*. 2014;93:1045-1053.
5. Albandar JM, Susin C, Hughes FJ. Manifestations of systemic diseases and conditions that affect the periodontal attachment apparatus: case definitions and diagnostic considerations. *J Periodontol*. 2018;89(Suppl 1):S183-S203.
6. Genco RJ, Borgnakke WS. Risk factors for periodontal disease. *Periodontol 2000*. 2013;62:59-94.
7. Borgnakke WS. Modifiable risk factors for periodontitis and diabetes. *Curr Oral Health Rep*. 2016;3:254-269.
8. Borgnakke WS. Non-modifiable" risk factors for periodontitis and diabetes. *Curr Oral Health Rep*. 2016;3:270-281.
9. Sutton JD, Salas Martinez ML, Gerkovich MM. Environmental tobacco smoke and periodontitis in United States non-smokers, 2009 to 2012. *J Periodontol*. 2017;88:565-574.
10. Suvan J, D'Aiuto F, Moles DR, Petrie A, Donos N. Association between overweight/obesity and periodontitis in adults. A systematic review. *Obes Rev*. 2011;12:e381-e404.
11. Borgnakke WS, Ylöstalo PV, Taylor GW, Genco RJ. Effect of periodontal disease on diabetes: systematic review of epidemiologic observational evidence. *J Periodontol*. 2013;84:S135-S152.
12. Preshaw PM, Alba AL, Herrera D, et al. Periodontitis and diabetes: a two-way relationship. *Diabetologia*. 2012;55:21-31.
13. Chapple IL, Van der Weijden F, Doerfer C, et al. Primary prevention of periodontitis: managing gingivitis. *J Clin Periodontol*. 2015;42(Suppl 16):S71-S76.
14. Sanz M, Marco Del Castillo A, Jepsen S, et al. Periodontitis and cardiovascular diseases: consensus report. *J Clin Periodontol*. 2020;47:268-288.
15. U S Department of Health and Human Services Healthy People 2030. Available at: https://health.gov/healthypeople/objectives-and-data/browse-objectives/oral-conditions/reduce-proportion-adults-aged-45-years-and-over-moderate-and-severe-periodontitis-oh-06. Accessed November,5 2020.
16. Dumancas GG, Adrianto I, Bello G, Dozmorov M. Current developments in machine learning techniques in biological data mining. *Bioinform Biol Insights*. 2017:11.
17. Batarseh F, Kulkarni A. Context-driven data mining through bias removal and data incompleteness mitigation. 1st workshop on evaluation and experimental design in data mining and machine learning. *Society for Industrial and Applied Mathematics*. 2019;2436:24-30.
18. Weir MR, Maibach EW, Bakris GL, et al. Implications of a healthy lifestyle and medication analysis for improving hypertension control. *Arch Intern Med*. 2000;160:481-490.
19. Learning feature representations with K-Means. In Coates A, Ng AY, Montavon G, Orr GB, Müller KR, eds. *Learning feature representations with K-Means. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2012:8_30.
20. Declaration of Helsinki. Recommendations guiding medical doctors in biomedical research involving human subjects. *Med J Aust*. 1976;1:206-207.
21. World Medical Association, World Medical Association. World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310:2191-2194.
22. Centers for Disease Control and Prevention, National Center for Health Statistics. National Health and Nutrition Examination Survey: NCHS Research Ethics Review Board (ERB) approval – NHANES 2009–2010, NHANES 2011–2012, NHANES 2013–2014. Available at: https://www.cdc.gov/nchs/nhanes/irba98.Htm. Accessed June,13 2020.
23. The PLOS Medicine Editors, The PLOS Medicine Editors. Observational studies: getting clear about transparency. *PLoS Med*. 2014;11:e1001711.
24. Eke PI, Page RC, Wei L, Thornton-Evans G, Genco RJ. Update of the case definitions for population-based surveillance of periodontitis. *J Periodontol*. 2012;83:1449-1454.
25. Huang Z. Extensions to the k-Means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*. 1998;2:283-304.
26. American Diabetes Association, American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2020. *Diabetes Care*. 2020;43:S14-S31.

27. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017-2018. *NCHS Data Brief*. 2020:1-8.
28. Centers for Disease Control and Prevention (CDC). Diabetes report card 2014. Atlanta, GA: CDC, U. S. Department of Health and Human Services. Available at: http://www.Cdc.Gov/diabetes/pdfs/library/diabetesreportcard2014.Pdf. Accessed June 13, 2020.
29. Feldman AL, Griffin SJ, Fharm E, et al. Screening for type 2 diabetes: do screen-detected cases fare better?. *Diabetologia*. 2017;60:2200-2209.
30. Estrich CG, Araujo MWB, Lipman RD. Prediabetes and diabetes screening in dental care settings: nHANES 2013 to 2016. *JDR Clin Trans Res*. 2019;4:76-85.
31. Borrell LN, Burt BA, Neighbors HW, Taylor GW. Social factors and periodontitis in an older population. *Am J Public Health*. 2004;94:748-754.
32. Borrell LN, Crawford ND. Socioeconomic position indicators and periodontitis: examining the evidence. *Periodontol 2000*. 2012;58:69-83.
33. Berge JM, Fertig A, Tate A, Trofholz A, Neumark-Sztainer D. Who is meeting the healthy people 2020 objectives? Comparisons between racially/ethnically diverse and immigrant children and adults. *Fam Syst Health*. 2018;36:451-470.
34. Sabbah W, Tsakos G, Chandola T, Sheiham A, Watt RG. Social gradients in oral and general health. *J Dent Res*. 2007;86:992-996.
35. Winning L, Polyzois I, Sanmartin Berglund J, Renvert S. Periodontitis and airflow limitation in older Swedish individuals. *J Clin Periodontol*. 2020;47:715-725.
36. Eke PI, Dye BA, Wei L, et al. Update on prevalence of periodontitis in adults in the United States: nHANES 2009 to 2012. *J Periodontol*. 2015;86:611-622.
37. Garcia D, Tarima S, Okunseri C. Periodontitis and glycemic control in diabetes: nHANES 2009 to 2012. *J Periodontol*. 2015;86:499-506.
38. Billings M, Holtfreter B, Papapanou PN, Mitnik GL, Kocher T, Dye BA. Age-dependent distribution of periodontitis in two countries: findings from NHANES 2009 to 2014 and SHIP-TREND 2008 to 2012. *J Clin Periodontol*. 2018;45(Suppl 20):S130-S148.
39. Beck JD, Philips K, Moss K, Divaris K, Morelli T, Offenbacher S. Advances in precision oral health. *Periodontol 2000*. 2020;82:268-285.
40. Eke PI, Thornton-Evans GO, Wei L, Borgnakke WS, Dye BA. Accuracy of NHANES periodontal examination protocols. *J Dent Res*. 2010;89:1208-1213.
41. Kingman A, Albandar JM. Methodological aspects of epidemiological studies of periodontal diseases. *Periodontol 2000*. 2002;29:11-30.


## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Ghassib IH, Batarseh FA, Wang HL, Borgnakke WS. Clustering by periodontitis-associated factors: A novel application to NHANES data. *J Periodontol*. 2020;92:1136-1150. https://doi.org/10.1002/JPER.20-0489