ISSN 1870-4069

An Analysis of Dietary and Demographic Data in Oral Health, Data from the National Health and Nutrition Examination Survey: A Preliminary Study

Nubia M. Chávez-Lamas[1], Laura A. Zanella-Calzada[2],
Carlos E. Galván-Tejada[2]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Odontología,
Clínica Comunitaria de Tacoaleche, Zacatecas, Zacatecas,
Mexico

[2] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica,
Zacatecas, Zacatecas,
Mexico

{lzanellac, ericgalvan}@uaz.edu.mx

Abstract. Dietary and demographics features has an influence in the general health status of the population in the world. Therefore in this paper is proposed an univariate analysis of 7 dietary and demographic features to study the impact on the oral status in order recognize the different determinants that contribute to modify in a negative way the oral health status. Univariate analysis is carried on applying an multi objective linear regression and evaluated in terms of area under the curve (AUC), p-value, sensitivity and specificity. Additionally, a multivariate model is done to evaluate confounder to increase the AUC. Preliminary results shows that individually water source is an important feature that affects oral health status.

Keywords: oral health, health status, linear regression, statistical analysis, univariate model, multivariate model.

1 Introduction

The World Health Organization (WHO), defined health as a physical, mental and social healthy status, not only the absence of diseases. This definition as evolved from that conceptual concept to a serie of quantitative scales that allows be measured of the general health status, therefor in 1994, the WHO propose the concept of quality of life as an individual perception of his life position, inside the context of cultural environment and the relationship with their objectives, expectations, standards and concerns. This new aspects that comprise the definition of health implies that behavior patterns, expectations and cultural are independent for each group, consequently, to evaluate this, three dimensions are proposed: General (lifestyle), particular (life conditions) and singular (type of life) [18,13].

Oral health is included in the general dimension, being an essential component to life quality of individuals, hence, nowadays oral health is a determining factor in individuals general health and of communities. For that reason, several studies about which characteristics of life, diet, demographic, social and others influence in the oral health are becoming an interesting niche study [18,13,7,15]. However, these studies can be influenced by several characteristics from the particular individuals, as was mentioned before, therefore, this features to evaluate oral health implies that varies depending on the condition of the country, genetic heritage, even political and public health services for each country [18].

These studies are done using a methodological scientific (survey), which is useful to prioritize, identify and solve oral health, moreover allows to generate prediction models that helps to decrease the incidence and prevalence of oral diseases. Is well-known that one of the main illness in oral health that affects 90 percent of the world population is dental caries and periodontal illness in a 80 percent [19], besides those are concentrated mainly in communities less favored by what are considered as oral health problems [15]. In last three decades, researchers develop different purpose surveys to evaluate life quality and the relationship with oral health, for instance, Social Impacts Of Dentals Disease, Geriatric Oral Health Assessment index, Dental Impact, Dental Impact on Daily Living, Oral Health Impact Profile, Oral Impacts on Daily Performances [9].

Through these instruments it is possible to recognize the different determinants that contribute to modify in a negative way the oral health status since its appearance depends on the conjugation of biological factors such as dental anatomy, diet where it is widely demonstrated the relationship between consumption and the appearance of oral diseases, both by historical evidence, observational, clinical studies and experimentation; socio-economic level, area of residence, educational level, occupation, housing characteristics, income, opportunities for general education, health, dental care as well as age and sex [17,4,14,8,19,6]

There are many risk factors and determinants of importance related to oral diseases and it is evident that the greater the degree of exposure to risk, the greater the probability of contracting or developing a condition [8], hence the importance of considering the present analysis that aims to determine some of the socioeconomic and dietary characteristics that directly influence oral health status. These may be considered as parameters that provide information on the normal or pathological state of an individual at a given time, a risk group and a specific place, in addition to helping to understand the oral health-disease process, in addition to establishing specific primary health care measures such as promotion, prevention, diagnosis, treatment and rehabilitation in a timely manner [2].

Therefore, the main contribution of this paper is to analyze as univariate models, individual demographic and dietary characteristics and the relationship of these with a general oral health status, as well as how an interaction of these characteristics (as a multivariate model) with the general status.

The analysis is carried on using a multi objective logistic regression and evaluated using a Receiver Operating Characteristic (ROC) curve, which allows us to study the sensitivity and specificity of each characteristic, just as the model comprised by all the selected characteristics in order to explain the relationship between demographic and dietary features with the oral health status.

This paper is organized as follows, in section 2 is presented a description of the data set used for this research and methods to carry on the study. In section 3 the experimentation conditions are presented. Results from univariate and multivariate analysis are shown in section 4. Finally conclusions and future work is described in section 5.

2 Materials and Methods

In this section is described the data set of National Health and Nutrition Examination Survey (NHANES), patients selection and the methods used to carry on the univariate analysis.

2.1 Data set Description

The NHANES is a national program that design studies to perform a survey to assess the health and nutritional status of adults and children in the United States, including all ethnic groups. The survey combines interviews and physical examinations, allowing to develop studies using clinical, para-clinical and demographic characteristics (features) of individuals. NHANES is a program founded by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

Content Description NHANES survey include several types of interviews, to cover a wide range of features, including demographic, socioeconomic, dietary, and health-related questions, described in detail in 1. One examination component that is critical to this study is dental care examination, which is carried on by trained medical personnel.

Table 1. NHANES data description.

Questionnaire Type	Description
Demographics	The demographics file provides individual, family, and household level information.
Examinations	Public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, evaluation of Federal health programs.
Dietary	Total nutrient intake.
Laboratory	Laboratory tests, which includes Cholesterol, Fasting Questionnaire, Hepatitis Tests, HIV, Urinary tests.
Questionnaire	information on: Acculturation, Alcohol Use, Health Insurance, Income.
Medication	Past 30 days, used or taken medication.

Meta Data Health and Nutrition Examination Surveys are comprised by interviews applied to 27,631 persons. In Table 2 are described in detail the number of persons and the features/parameters of the groups included in the sample.

Table 2. Patient demographic information.

Characteristic	**NHANES**
Age of civilian	All ages from birth
Geographic areas	Unaited States
Average number of sample persons per household	2
Number of study locations	60
Domains for oversampling	Predesignated: 87 subdomains of sex-age groups for non-Hispanic black persons, non-Hispanic non-black Asian persons, and Hispanic persons. Oversampled: Hispanic persons, non-Hispanic black persons, non-Hispanic non-black.
Number of selected persons	27,631
Number of interviewed persons	20,491
Number of examined persons	19,644

2.2 Data Analysis

In this work, as mentioned before, were conducted an univariate and multivariate searches. The univariate search was conducted using the demographics and dietary features being subjected to a statistical analysis; while the multivariate search was conducted using the complete feature set, including the examinations features.

The statistical analysis consisted of submitting each of the features to a linear regression to obtain the univariate models; then, all features together developed a multivariate model trough a linear regression too.

Linear regression is an analysis which consists on a statistical technique for modeling the relationship between features. The simplest representation of a model obtained by this method is represented in Equation 1, where y is the dependent variable or the outcome feature, β_0 is the intercept, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of terms needed. Finally, the difference between the real outcome and the outcome proposed by the model is the error of the model, ϵ; this variable is known as a statistical error due to it's a random variable that measures the model failure for fitting the data exactly [16]:

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1}$$

The statistical validation consisted on obtaining the P-value, the odds ratio and the area under the receiver operating characteristic (ROC) curve, known as the AUC.

The P-value or the observance significance level is the smallest value obtained where the null hypothesis can be rejected [13]; while the AUC is a standard method to evaluate the accuracy of the classification model [12], finally, the odds ratio represents the ratio of the probability that an event of interest occurs against the probability that the same event doesn't occur [3].

All the data analysis were realized with the free software *R* (version 3.3.1) [20] and its packages, *pROC* (Version 1.8, 2015-06-10) [21], *epitools* (Version 0.5-7, 2012-09-30) [1], *ResourceSelection* (Version 0.2-6, 2016-02-15) [10] and *randomForest* (version 4.6-12, 2011-10-18) [11].

3 Experiments

The process realized in the univariate and multivariate models experimentation and their statistical analysis is described in this section. In Figure 1 is presented a flowchart of the followed methodology.

Firstly, the NHANES data analysis and the features selection for this research were realized by an expert dentist, according to the information obtained from the literature. Then, a new dataset with the extracted features was obtained (Figure 1 A)). Followed by a data preprocessing, where is initially described the imputation of missing data (Figure 1 B)) and then, for the univariate analysis, it was necessary removing the features that are unable to contribute meaningful information due to their contents. The univariate models development and models validation were realized in base of an statistical process (Figure 1 C)).

Finally, for the multivariate model development all features were taking into account and the model validation is also carried out (Figure 1 D)). The validation process is performed with the intention of evaluating the contributions of the results.

3.1 Dataset Preprocessing

The dataset used for this research was mostly contained by demographic and examination features, also a dietary feature was present [5] .

- Demographic features:
  - DMDMARTL: Describes the marital status,
  - INDFMIN2: Describes the annual family income,

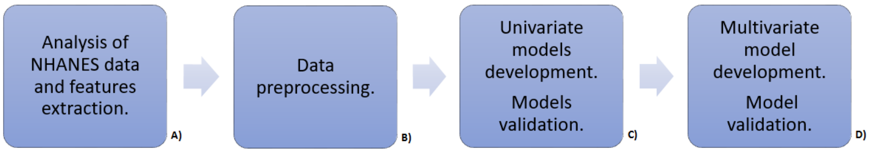
Analysis of NHANES data and features extraction.
A)
Data preprocessing.
B)
Univariate models development.
Models validation.
C)
Multivariate model development.
Model validation.
D)

Fig. 1. Flowchart of the methodology followed. A) Dataset analysis and features extraction, B) Univariate models development and their statistical validation, C) Multivariate model development and its statistical validation.

  - RIAGENDR: Describes the gender of the participant,
  - DMDHHSIZ: Describes the total number of people in the household,
  - DMDEDUC3: Describes the education level on youths from 6 to 19 years,
  - DMDEDUC2: Describes the education level on adults from 20 years old.
- Dietary feature:
  - DR1TWS: Describes the tap water source.
- Outcome feature:
  - OHDEXSTS: Describes the overall oral health exam status (from 1 being complete to 3 being not done).

Examination features were initially removed for the univariate analysis, since they contain information about the health center where the patient was treated and each health center is represented by a different feature, therefore, each feature presents a large amount of missing data, becoming impossible to perform an univariate analysis for them.

The remaining features contained some missing values represented as Not a Number (NaN). These missing values were imputed through the $rfimpute$ function from $randomForest$ package, consisting on the substitution of NaN's for the value of the median of the column where the features are found.

3.2 Univariate Models Development and Models Validation

The development of the univariate models was realized by a linear regression process between each of the features and the outcome feature, which was related to the general status of oral health.

After the univariate models were obtained, they were subjected to an univariate statistical analysis, obtaining their P-values, odds ratios, AUC and ROC curves.

3.3 Multivariate Model Development and Model Validation

For the multivariate model development, which was realized by a linear regression between the feature set and the outcome feature, was also included the examination features that weren't used for the univariate analysis. The information contained in the examination features was all joined in only one feature, avoiding problems in the model development because of the missing data, taking into consideration that for each patient these features only contained information in one of them, because they were related to the health center where the patients were attended and most of them were attended only in one.

Finally, for the multivariate model validation were obtained the P-value, AUC and ROC curve of the model. Odds ratio wasn't calculated for this model because this parameter is obtained for specific features and not for a set of them.

4 Results

Results obtained from the univariate statistical analysis are presented in Table 3, which is contained by every feature of the dataset and its respective P-value, odds ratio and AUC, in ascendant order according to the AUC values. For this analysis, features related with the health center where the patients were treated (examination features), weren't taken into account.

Table 3. Univariate statistical analysis.

Feature	P-value	AUC	Odds ratio	2.5%	97.5%
DMDMARTL	0.683	0.496	0.999	0.994	1.003
INDFMIN2	0.326	0.501	0.999	0.998	1.000
RIAGENDR	0.299	0.502	0.999	0.999	1.000
DMDHHSIZ	0.585	0.510	1.001	0.996	1.006
DMDEDUC3	0.744	0.521	1.005	0.997	1.003
DMDEDUC2	0.472	0.521	1.003	0.993	1.013
DR1TWS	0.101	0.540	1.000	0.999	0.101

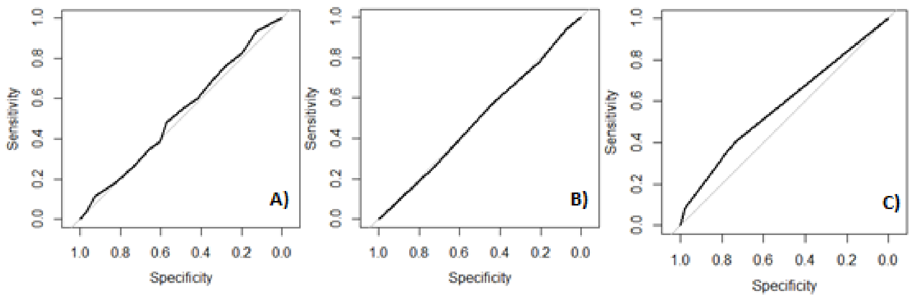

Fig. 2. ROC curves obtained from the most significant features in the univariate analysis, A) DMDEDUC3, B) DMDEDUC2, C) DR1TWS.

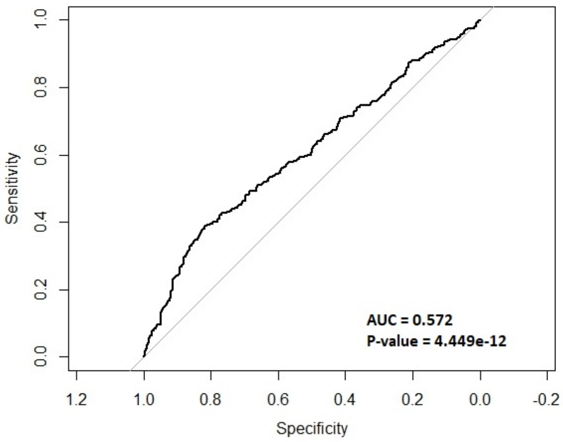

Fig. 3. ROC curve obtained from the multivariate analysis (AUC = 0.572, P-value = 4.449e-12).

From the univariate analysis results, the three most representative univariate models, according to their AUC values, present their ROC curves in Figure 2. In Figure 2 A) is shown the ROC curve of DMDEDUC3 feature, in Figure 2 B) is shown the ROC curve of DMDEDUC2 feature and in Figure 2, C) is shown the ROC curve of DR1TWS feature.

In multivariate statistical analysis, the ROC curve generated is shown in Figure 3, obtaining a P-value of 4.449e-12 and a value of AUC of 0.572. For this analysis, features related to the health center where the patients were treated (examination features) were taken into account.

5 Conclusion and Future Work

It is well known that the health problems that arouse the greatest interest are those that represent a risk of death or permanent disability, and carry with them the doubt as to the possibility of attacking a certain person.

Commonly, oral health problems do not arouse the spontaneous interest of the community, which is why identifying the determinants through such analysis in the population allows informing, educating and motivating appropriate oral health care because it depends on whether the population shows interest in receiving different treatments and thus improve oral health.

This analysis has shown that around a disease there are many factors that cause or aggravate it. When referring to the issue of socio-cultural health determinants (SHD) has become a latent concern for health professionals and oral health professionals since most of the problems derive from inequality or inequity in health and are linked to other determinants which produce different effects in each risk group with respect to their conditions and lifestyle, generated in the short, medium or long term.

Demographic features that were used for this research showed that none of them can significantly predict the health status of the population in an univariate approach, this can be observed at the statistical analysis where all of the P-values were > 0.06, which is taken as the standard value to consider a feature to be significant; also, the AUC values and the ROC curves showed a similar result, since the higher AUC is 0.540, which means that the true positives / true negatives proportion is 54% for this specific feature, DR1TWS (dietary feature). Odds ratios didn't show a higher probability than 1 for any feature in relationship with a health condition. Although the predictive capacity of this feature is better than a blind test, its contribution isn't so statically significant.

Multivariate model showed better statistical results than univariate models. In this approach, where the examination or health centers features were also used, the P-value obtained, 4.449e-12, is statistically highly significant, which represents that the alternative hypothesis has a high probability of being true, it means that the model has an influence on the health condition. The AUC and the ROC curve presented a true positives / true negatives proportion of 57.2%, which is higher than the AUC values of the univariate models. The improvement of statistical values in the multivariate model allows to conclude that the demographic and dietary features can help to evaluate the health status in combination with the examination features. By this, it's possible to consider that social security, economic income, type of health service and other similar factors may help to determine the patient's health status, specifically the oral health, according to the data used for this work.

As future work is proposed add more dietary and medical condition features, but including a clever feature selection than can lead to a high AUC model cleaning features that can decrease specificity and sensitivity of a oral health prediction model, in addition, complex techniques of machine learning, as neural networks, elastic networks or similar can be used to tackle this problem.

References

1. Aragon, T.: EpiTools: Epidemiology Tools. R package version 0.5-7 (2016)
2. Arango, V., Sandra, S.: Biomarcadores para la evaluación de riesgo en la salud humana. Revista Facultad Nacional de Salud Pública 30(1), 75–82 (2012)
3. Bland, J.M., Altman, D.G.: The odds ratio. Bmj 320(7247), 1468 (2000)
4. Castañeda Abascal, I.E., Lok Castañeda, A., Molina, L., Manuel, J.: Prevalencia y factores pronósticos de caries dental en la población de 15 a 19 años. Revista Cubana de Estomatología 52, 21–29 (2015)
5. for Disease Control, C., Prevention, et al.: National health and nutrition examination survey, 2011-2012 (2013)
6. Escalona, T.P., Ortiz, H.R.C., Palomino, Y.P., Tamayo, M.I., Rodríguez, M.I.R.: 08-relación entre factores de riesgos y caries dental relationship between risk factors and dental caries. MULTIMED Revista Médica Granma 19(4) (2017)

7. Espinosa González, L.: Cambios del modo y estilo de vida; su influencia en el proceso salud-enfermedad. Revista Cubana de Estomatología 41(3), 0–0 (2004)
8. Gispert Abreu, E.d.l.Á., Castell-Florit Serrate, P., Herrera Nordet, M.: Salud bucal poblacional y su producción intersectorial. Revista Cubana de Estomatología 52, 62–67 (2015)
9. González, C.F., Franz, L.N., Sanzana, N.D.: Determinantes de salud oral en población de 12 años. Revista clínica de periodoncia, implantología y rehabilitación oral 4(3), 117–121 (2011)
10. Lele, S.R., Keim, J.L., Solymos, P., Solymos, M.P.: Package ResourceSelection (2017)
11. Liaw, A., Wiener, M.: The randomforest package. R News 2(3), 18–22 (2002)
12. Lobo, J.M., Jiménez-Valverde, A., Real, R.: Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17(2), 145–151 (2008)
13. Martínez Abreu, J., Capote Femenias, J., Bermúdez Ferrer, G., Martínez García, Y.: Determinantes sociales del estado de salud oral en el contexto actual. MediSur 12(4), 562–569 (2014)
14. Martínez Abreu, J., Castell-Florit Serrate, P., Llanes Llanes, E., Morales Aguiar, D.R., Sánchez Barrera, O., et al.: Componente bucal y determinantes sociales en el análisis de la situación de salud. Revista Cubana de Estomatología 52, 53–61 (2015)
15. Medina Solís, C.E.: Políticas de salud bucal en México: disminuir las principales enfermedades. Una descripción (2006)
16. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. John Wiley & Sons (2015)
17. Narváez Chávez, A.M.: Asociación entre el conocimiento de los padres sobre salud bucal y uso de técnicas educativas con relación a la presencia de biofilm y caries en infantes. Master's thesis, Quito: UCE (2017)
18. Ojeda-Garcés, J.C., Oviedo-García, E., Salas, C.E.S.: Streptococcus mutans y caries dental. Odontologica 26(1) (2013)
19. Ospina, D., Herrera, Y., Betancur, J., Agudelo, H.B., López, A.P.: Higiene bucal en la población de san francisco antioquia y sus factores relacionados. Revista Nacional de Odontología 12(22), 23–30 (2016)
20. Ripley, B.D.: The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network 1(1), 23–25 (2001)
21. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 12(1), 77 (2011)