HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

Stud Health Technol Inform. 2018 ; 251: 253–256.

Machine Learning to Identify Behavioral Determinants of Oral Health in Inner City Older Hispanic Adults

Sunmoo YOON, RN PhD[a], **Thomas CHOI, DDS**[b], **Michelle ODLUM, EdD, MPH**[c,†], **Dennis A. MITCHELL, DDS, MPH**[b], **Ian M. KRONISH, MD, MPH**[a], **Karina W. DAVIDSON, PhD**[a], and **Joseph FINKELSTEIN, MD, PhD**[b]

[a]Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[b]Department of Medicine, College of Dental Medicine, Columbia University, New York, NY, USA

[c]School of Nursing, Columbia University, New York, NY, USA

Abstract

We applied machine learning techniques to a community-based behavioral dataset to build prediction models to gain insights about minority dental health and population aging as the foundation for future interventions for urban Hispanics. Our application of machine learning techniques identified emotional and systemic factors such as chronic stress and health literacy as the strongest predictors of self-reported dental health among hundreds of possible variables. Application of machine learning algorithms was useful to build prediction models to gain insights about dental health and minority population aging.

Keywords

Dental health; population aging; Hispanics; deep learning

1. Introduction

Dental conditions disrupt quality of life among older adults. Aging is associated with recession of gum, loss of tooth, root caries, dry mouth, periodontitis and gingivitis leading to pain and inadequate nutrition.[1] Hispanics comprise a fast growing ethnic group yet have one of the poorest dental health conditions among any racial or ethnic groups in the U.S.[2] While Healthy People 2020 aspires to the elimination of dental health disparities, limited access to dental care persists within the U.S, resulting in profound, continuing dental health disparities.[1, 2] In this study, we applied machine learning techniques to explore variables that are associated with urban Hispanics' dental health among older adults as a foundation for future targeted intervention development.

† Michelle Odlum, EdD, Columbia University School of Nursing, 630W 168 street, mail code 6, New York, NY, 10032, USA, mlo12@cumc.columbia.edu.

The Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER)[3] project built an informatics infrastructure to understand health behaviors to improve the health of an urban underserved population dwelling in northern Manhattan in New York City. The WICER dataset is available to investigators within the institution, and those outside of the institution with data use agreement and institutional review board approval. One component of the infrastructure is a survey which was collected by bilingual community health workers. The WICER survey dataset contains 925 variables from 5429 Hispanics including physiological, environmental, behavioral, patient-reported outcomes, and sociodemographic factors.[3]

2. Methods

Machine learning techniques offer power to search numerous possible relationships and efficiently remove redundant variables among hundreds of variables. The socio-ecological framework and the data mining process model[3] guided our analytic process. The data mining process includes: 1) understanding the problem, 2) understanding data, 3) preprocessing data, 4) reducing dimensionality, 5) applying mining algorithms, and 6) interpreting results. We applied machine learning techniques to the WICER dataset to search and evaluate attributes among 925 variables and to build a risk behavioral model. We used Weka 3.7.12, a collection of machine learning algorithms and BigML for our analytic process.

We extracted 925 variables for 5429 Hispanics from the local REDCap database and queried 2007 records of Hispanics older adults (age ≥55). During the prime filtering, 80 of total 925 variables including demographics, predisposing and enabling factors (e.g., The Newest Vital Sign score for health literacy) were selected as relevant based on the literature.[4] During feature selection, we applied six machine learning algorithms[5] to remove redundant variables and select variables that are strongly correlated to the dichotomous dependent variable (‘good self-rated teeth and gum health’ versus ‘poor self-rated teeth and gum health’). Six machine learning algorithms with default configuration in Weka and a deep learning algorithm (deep neural network) in BigML were chosen to avoid algorithm dependency because selected features can vary by machine learning algorithm.[5] We iteratively selected nine final variables based on the criteria of clinical meaningfulness and identification in multiple of the six algorithms. We then organized the variables into four conceptual categories: emotional, behavioral, systemic (issues inherent in the overall socio-structural system rather than due to individual factor), and environmental factors.

Next, we iteratively applied ten data mining algorithms in Weka (J48, ADTree, DecisionStump, RandomForest, BayesNet, SMO, AdaBoost M1, Bagging, PART, Random Tree)[5] using the top features to build the prediction models for self-reported dental health among Hispanic older adults. As with feature selection, we used multiple algorithms to avoid algorithm dependency. For 10-fold cross-validation, the WICER dataset was randomly divided into training and evaluating datasets for the model validation before applying the algorithm. We chose the final models based on predictive accuracy (i.e., correctly classified survey participants), the area under the receiver operating characteristic curve (AUC), and the model interpretability. Last, we interpreted the models according to clinical meaningfulness and applicability.

3. Results

Study participants (n=2,007), age 55–100 (mean 65.3 ± 7.8) were predominantly female (n=1,484, 73.9%) and Spanish speaking (n=1,222, 60.9%), an education level of eighth grade of less (n=1,045, 52.1%), and with a poor health literacy score. One third (n=683, 34.0%) of the participants reported that they perceived themselves as having 'poor teeth and gum health.' Table 1 summarizes descriptive statistics for study variables. Health literacy score, depressive and anxiety symptoms, the availability of large selection of fruits and vegetables in neighborhood (environmental factor) were the variables selected by at least five machine learning algorithms (Figure 1). While self-reported general health was depicted as a strong factor by four algorithms, it was excluded in the final prediction model because the concept of self-reported general health likely overlaps with self-reported oral health.

This study mainly found that among the participants with low perceived stress and limited health literacy (n=523, 26%), the prediction model in Figure 1 shows that individuals experiencing anxiety symptoms more than 2 days in last month were more likely to report 'poor dental health', while individuals experiencing less frequent anxiety symptoms (<=2 days last month) were more likely to report 'good dental health' if large selection of fruits and vegetables are available in their neighborhood.

4. Discussion and Conclusion

Among many risk factors, we found that emotional factors such as anxiety and depressive symptoms (red) and health literacy (green) were more strongly associated with self-reported dental health compared to other demographic or physiological factors among Hispanic older adults living in New York City. This is consistent with previous epidemiological findings among other ethnic/racial minorities in New York City,[1] showing that low health literacy can adversely influence dental health outcomes.

A large body of literature reports on the association between compromised dental health literacy and poor dental health outcomes. This study adds to that body of literature by revealing that health literacy using a common Newest Vital Sign tool was a main predictor of poor dental health among Hispanic older adults, when evaluating hundreds of competing variables in our machine learning models. In this study (Figure 1), individuals who had the worst score in the health literacy, were less likely to report 'good dental health' regardless of the level of chronic stress. Yet, one of the greatest challenges facing dental medicine continues to be the lack of diversity in the providers; providers from diverse ethnic backgrounds with cultural competency may be best suited to bridge the health literacy gap for these patients as future work.[1]

Machine learning[5] was useful for efficiently removing redundant variables and for building prediction models for self-reported dental health from a large and complicated dataset with over 900 variables. Data visualizations (e.g., tree infographics) of the prediction model results were helpful to detect patterns and gain insights about risk factors for minority population dental health. Addressing health literacy through improved communication skills at the dental system level as well as through addressing inequities in education at the

systems level may reduce disparities in dental health. This study was conducted in a single city where predominantly represents limited ethnicities.

The machine learning including deep learning techniques revealed health literacy and emotional factors as the strong risk predictors of self-reported dental health among urban Hispanic older adults. This new knowledge adds insights about dental health and minority aging population for future intervention.

Acknowledgments:

U.S. federal grant WICER (R01HS019853, PI: Bakken).

References

[1]. Shelley D, Russell S, Parikh NS, Fahs M. Ethnic disparities in self-reported oral health status and access to care among older adults in nyc. J Urban Health. 88 (2011), 651–662 [PubMed: 21850607]

[2]. Patrick DL, Lee RS, Nucci M, Grembowski D, Jolles CZ, Milgrom P. Reducing oral health disparities: A focus on social and cultural determinants. BMC Oral Health. 6 (2006), Suppl 1:S4 [PubMed: 16934121]

[3]. Yoon S, Suero-Tejeda N, Bakken S. A data mining approach for examining predictors of physical activity among urban older adults. J Gerontol Nurs. 41 (2015), 14–20

[4]. Valencia A et al. Racial and ethnic disparities in utilization of dental services among children in iowa: The latino experience. Am J Public Health. 102 (2012), 2352–2359 [PubMed: 22698039]

[5]. Beam AL, Kohane IS. Big data and machine learning in health care. Jama. 319 (2018), 13–1317–8.

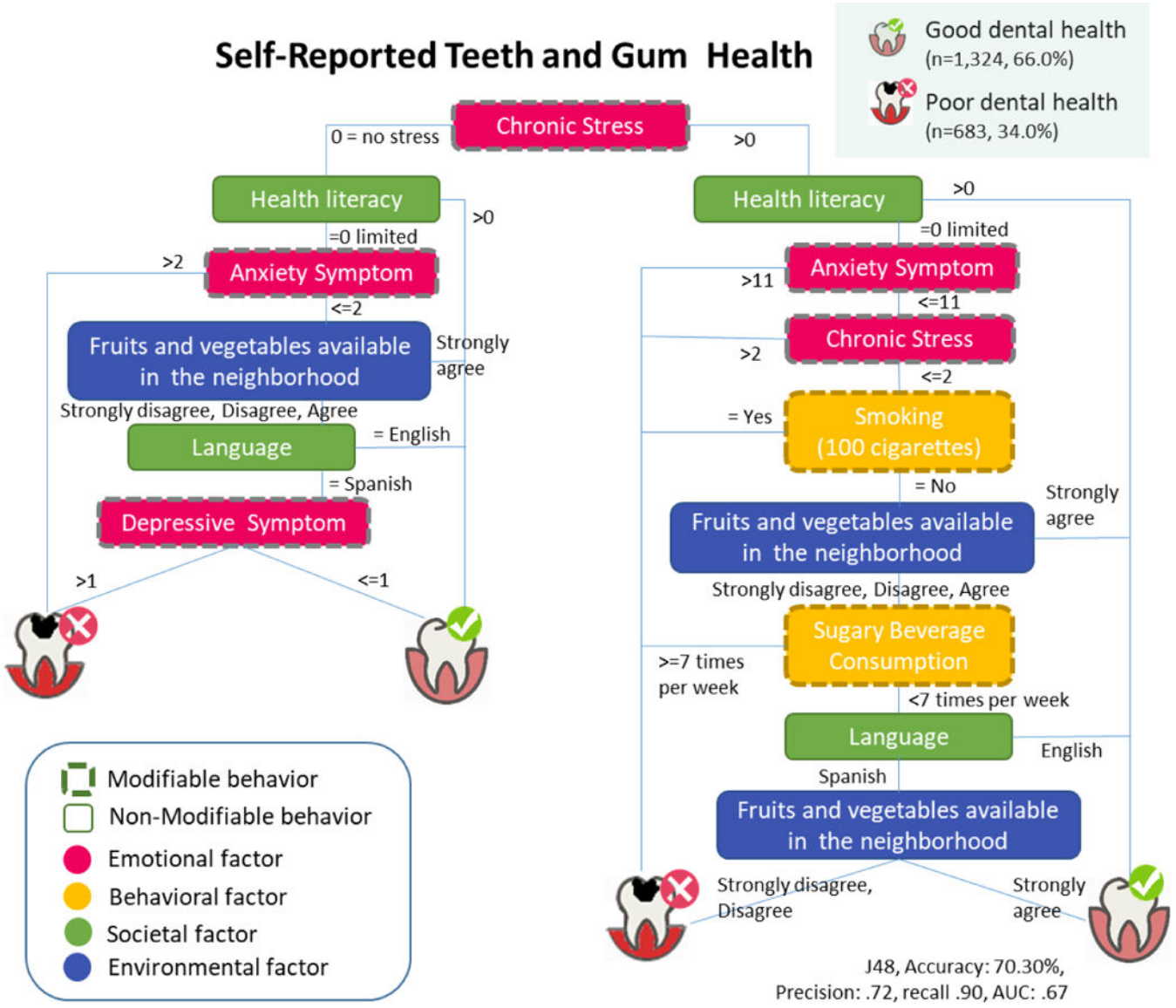

Figure 1.
Prediction models for self-reported Teeth and Gum Health

Table 1.

Descriptive Statistics for Study Variables (n=2,007)

Variables*	Mean (SD), N (%)
Emotional factor	
Perceived Stress[a] [0–5, 5: worst]	0.6 (SD 1.0)
Anxiety symptoms[b]- days of having anxiety symptoms last month [0–30 days, 30: worst]	3.5 (SD 7.4)
Depressive symptoms[c] [0–27, 27:worst]	2.2 (SD 4.6)
Behavioral factor	
Smoking (100 cigarettes) in life	N=263 (13.1%)
Sugary beverage consumption [0–70 times/week]	2.9 (SD 4.3)
Societal factor	
Health literacy[d] [0–6, 0:limited health literacy]	1.7 (SD 1.8)
Environmental factor	
Fruits and vegetable available in the neighborhood Strongly agree 746, (37.5%), Agree 960 (48.3%), Disagree 239 (12.0%), Strongly disagree 44 (2.2%)	

[a]Perceived Stress Scale (PSS) (Cohen, Kamarck, & Mermelstein, 1983)

[b]Centers for Disease Control and Prevention Health-related quality of life (CDC HRQOL-14)

[c]Modified Patient Heath Questionnaire-9 (modified PHQ-9 depression)

[d]Newest Vital Sign-Health literacy (Weiss et al., 2005)

*missing data ≤1% except for chronic stress (3%), depressive symptom (4%) and anxiety symptom (5%)