MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-1281

Identifying Patients' Smoking Status from Electronic Dental Records Data

Jay Patel[a,b], Zasim Siddiqui[a], Anand Krishnan[a], Thankam Thyvalikakath[a,c]

[a] *Indiana University School of Dentistry, Indianapolis, IN, USA*
[b] *Department of Bio-Health Informatics, Indiana University School of Informatics and Computing, Indianapolis, IN, USA*
[c] *Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA*

Abstract

Smoking is a significant risk factor for initiation and progression of oral diseases. A patient's current smoking status and tobacco dependency can aid clinical decision making and treatment planning. The free-text nature of this data limits accessibility causing obstacles during the time of care and research utility. No studies exist on extracting patient's smoking status automatically from the Electronic Dental Record. This study reports the development and evaluation of an NLP system for this purpose.

Keywords:
Smoking; Dental Records; Clinical Decision-Making.

Introduction

Smoking is a significant risk factor for initiation and progression of oral diseases such as periodontal disease, dental caries, and oral cancers. Therefore, it is crucial for dental clinicians to be aware of patient's current smoking status and tobacco dependency [1]. This can help clinicians to make decisions and to take preventive measures as well as for treatment planning. In addition, this information can be used for conducting large scale research studies such as studies related to the association and correlation of smoking and oral diseases. However, due to free-text nature of the data, the access to this information can cause obstacles during the time of care and can be limiting for research purposes [2]. Historically, manual review is required in order to use this information. However, manual chart review can be labor intensive, expensive, and time-consuming. Natural language processing (NLP) can automatically extract patients' smoking status from these histories to reduce human effort. Currently, no study exists on extracting patients' smoking status automatically from Electronic Dental Record (EDR) data. This study reports on developing and evaluating an NLP system to identify patient's smoking status from free text data in EDR.

Methods

This study was approved by the Indiana University Institutional Review Board (Study #:1611054551) and conducted at the Indiana University School of Dentistry (IUSD). We extracted de-identified clinical notes of patients who underwent oral examination at IUSD between December 31, 2011-January 1, 2012 from the EDR. Next, two clinicians trained in informatics manually reviewed and annotated 555 sentences describing patients' smoking status as a smoker, non-smoker, or past smoker. We performed Cohen's kappa statistical test to find the inter-annotation agreement (IAA) between the two annotators and any disagreement was resolved through discussion and consensus. We considered this dataset our gold standard and we divided this dataset into training (389) and testing (166) sets. We used the training dataset to develop our NLP algorithm. We evaluated the performance of NLP algorithm by testing its accuracy, precision, and recall on the test set.

Results

We observed an IAA of 92%. Our system achieved high precision (98%), recall (98%), and f measure (99%) in differentiating smokers and non-smokers. We observed 80% precision, recall, and accuracy when classifying patients into past smokers. The reason behind the moderate performance could be due to variations in documentations concerning past smoking status of patients in the EDR.

Conclusion

Our NLP system performed excellent in classifying patients' smoking status into smoker and non-smoker and had moderate performance in classifying past smokers. In the future, we will annotate more smoking histories and increase the size of our training set. In addition, we will capture more variations in documenting past smokers. We will run this algorithm on bigger dataset and use this information for research purposes.

References

[1] G., Smoking, smokeless tobacco, and alcohol consumption as contributing factors to periodontal disease, *Northwest Dent J* **95** (2016): 37-42.

[2] DT. Heinze, ML. Morsch, BC. Potter, RE. Sheffer Jr, Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology, *J Am Med Inform Med* **15**(2008.): 40-43.

Address for correspondence

Thankam Paul Thyvalikakath, DMD, MDS, PhD
Associate Professor & Director of Dental Informatics Core;
Research Scientist, Regenstrief Institute, Inc.
Indiana University School of Dentistry, IUPUI
1121 W Michigan Street, Room DS314
Indianapolis, IN 46202
Email: tpt@iu.edu
Phone: (317) 274-5460.