Leveraging Electronic Dental Record Data to Classify Patients Based on Their Smoking Intensity

J. Patel[1,2] Z. Siddiqui[1] A. Krishnan[1] T. P. Thyvalikakath[1,2,3]

[1] Dental Informatics Core Division, Department of Cariology, Operative Dentistry, and Dental Public Health, Indiana University School of Dentistry, Indiana University – Purdue University Indianapolis, Indianapolis, Indiana, United States
[2] Department of Bio-Health Informatics, School of Informatics and Computing, Indiana University – Purdue University Indianapolis, Indianapolis, Indiana, United States
[3] Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, United States

Address for correspondence Thankam Paul Thyvalikakath, DMD, MDS, PhD, Dental Informatics Core Division, Department of Cariology, Operative Dentistry & Dental Public Health, Indiana University School of Dentistry, 1050 Wishard Boulevard, R2206, Indianapolis, IN 46202, United States (e-mail: tpt@iu.edu).

Methods Inf Med

Abstract

Background Smoking is an established risk factor for oral diseases and, therefore, dental clinicians routinely assess and record their patients' detailed smoking status. Researchers have successfully extracted smoking history from electronic health records (EHRs) using text mining methods. However, they could not retrieve patients' smoking intensity due to its limited availability in the EHR. The presence of detailed smoking information in the electronic dental record (EDR) often under a separate section allows retrieving this information with less preprocessing.

Objective To determine patients' detailed smoking status based on smoking intensity from the EDR.

Methods First, the authors created a reference standard of 3,296 unique patients' smoking histories from the EDR that classified patients based on their smoking intensity. Next, they trained three machine learning classifiers (support vector machine, random forest, and naïve Bayes) using the training set (2,176) and evaluated performances on test set (1,120) using precision (P), recall (R), and F-measure (F). Finally, they applied the best classifier to classify smoking status from an additional 3,114 patients' smoking histories.

Results Support vector machine performed best to classify patients into smokers, nonsmokers, and unknowns (P, R, F: 98%); intermittent smoker (P: 95%, R: 98%, F: 96%); past smoker (P, R, F: 89%); light smoker (P, R, F: 87%); smokers with unknown intensity (P: 76%, R: 86%, F: 81%), and intermediate smoker (P: 90%, R: 88%, F: 89%). It performed moderately to differentiate heavy smokers (P: 90%, R: 44%, F: 60%). EDR could be a valuable source for obtaining patients' detailed smoking information.

Conclusion EDR data could serve as a valuable source for obtaining patients' detailed smoking information based on their smoking intensity that may not be readily available in the EHR.

Keywords

- electronic dental record
- smoking intensity
- information extraction
- electronic health record
- dental informatics
- machine learning classifiers

received
May 8, 2018
accepted after revision
October 11, 2018

DOI https://doi.org/10.1055/s-0038-1675817.
ISSN 0026-1270.

Background and Significance

Smoking is a major risk factor of common oral diseases such as dental caries, and periodontal disease,[1] which, if left untreated, leads to tooth loss and poor quality of life.[2–5] Moreover, people who smoke have an increased risk to develop oral cancers compared with nonsmokers.[6] As a result, dental clinicians routinely assess and record their patients' detailed smoking status such as smoking intensity and duration (cigarettes or packs smoked per day/year).[7,8] They use this information to assess their patients' risk of developing oral diseases, and the prognosis of treatment provided.[8,9] With the increased use of electronic dental record (EDR) to document patient care, clinical information including patients' detailed smoking histories are available electronically.[9,10] This offers the opportunity to utilize it for clinical care and research purposes.[10,11] Similar to electronic health record (EHR) data,[12–21] smoking histories are mostly documented as free-text (unstructured format) in the EDR and could be time-consuming to retrieve them manually. Natural language processing and text mining approaches have been successfully used to retrieve patients' smoking status from the EHR.[13–21] However, scant reports exist on the automated retrieval and classification of patients' smoking status from their EDR.

Most studies in medicine have classified patients' smoking status superficially based on their smoking statuses (past, current, nonsmoker, and unknowns).[13–17,21] These studies applied machine learning (ML) classifiers such as support vector machine (SVM), naïve Bayes, random forest, and decision tree to retrieve patients' smoking-related information from the EHR.[13–21] A few studies also applied traditional features such as unigrams, bigrams, and parts of speech tags in combination with rules.[15,17] Hybrid systems have also been developed utilizing a combination of topic modeling and SVM to classify patients' smoking status from unstructured EHR data.[21] However, these studies[13–17,21] did not classify patients based on their smoking intensity (cigarettes or packs smoked per day/year), although determining patients' smoking intensity is important to assess disease prognosis and treatment outcomes.[9] A recent study reported the lack of availability of smoking intensity in the EHR as a limitation to classify patients based on their smoking intensity .[15]

Since dental clinicians gather their patients' detailed smoking status, EDR data could be a rich resource to retrieve patients' detailed smoking information. In addition, this information is often documented in response to the question, "Do you use tobacco, or alcohol?" in the social history section and not in progress/clinical notes (that may contain chief complaint, treatment progress, etc.) as seen in the EHR.[15] In this study, we have considered only responses to this question that contains patients' smoking histories (hereon referred to as "smoking histories" in this article). This distinction in recording smoking history makes it easier to retrieve patients' smoking information from the EDR than from the EHR.

Objective

The objective of this study was to determine patients' detailed smoking status based on their smoking intensity from EDR utilizing ML classifiers. We applied three ML classifiers (SVM, naïve Bayes, and random forest) to extract this information due to following reasons: the presence of detailed smoking documentation and under a separate section in EDR made it easier to train the classifiers to detect a consistent pattern of smoking intensity information across patients' EDR.

Methods

Our approach consisted of the following steps. First, we retrieved 6,410 unique patient records. Second, we created guidelines to annotate patients' detailed smoking histories based on smoking intensity in the EDR. Next, we developed a reference standard consisting of manually annotated unique patient smoking histories from 3,296 records. We then randomized and split this reference standard into training and test sets. Subsequently, we trained three ML classifiers using the training set and evaluated their performances using the test set. We applied the classifier that performed best on the remaining 3,114 unique patients' smoking histories. From here on the data set (3,114 [49%]) used for extracting patients smoking information automatically will be referred as "new data set." We also evaluated the ML classifier's performance on this new data set using a subset of 315 randomly selected smoking histories (see ►Fig. 1). We describe each step in detail below.

Data Extraction and Preprocessing

Our data set consisted of 6,410 unique patients who underwent comprehensive oral evaluation from January 1, 2009, to December 31, 2011, at the Indiana University School of Dentistry. We included only patients' smoking histories from their single visits during this period. If we found multiple visits of these patients, we considered smoking histories reported during their latest dental visit. We extracted, preprocessed, and standardized the patients' smoking histories by removing all nonalphabetical and nonnumerical characters and converting all texts into lower case. We then converted each patient's smoking history into individual text files.

Annotation Guidelines and Reference Standard

As displayed in ►Table 1, we created guidelines based on the existing literature in medicine and dentistry[13,14,16–18,22–24] to annotate patients' smoking status including smoking intensity. Next, using these guidelines, two researchers independently annotated 3,296 unique patients' smoking histories. The application and evaluation of text mining methods such as supervised ML methods depend on validated manually annotated texts, which are often referred as the reference standard or gold standard corpus.[25] It is also used to train classifiers to learn patterns in written text when using supervised ML in text mining.[25] We used the

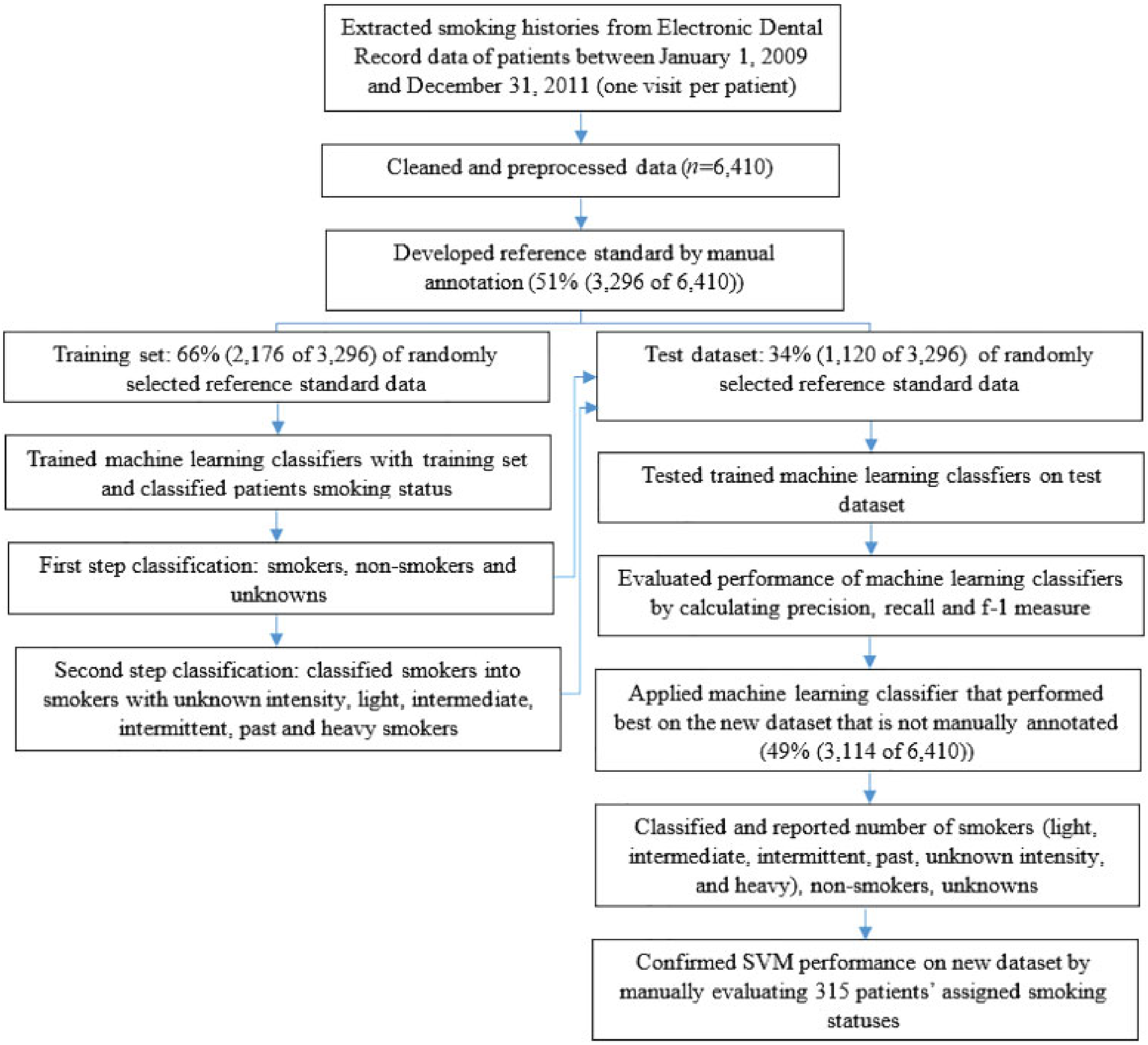
Extracted smoking histories from Electronic Dental Record data of patients between January 1, 2009 and December 31, 2011 (one visit per patient)
Cleaned and preprocessed data (n=6,410)
Developed reference standard by manual annotation (51% (3,296 of 6,410))
Training set: 66% (2,176 of 3,296) of randomly selected reference standard data
Trained machine learning classifiers with training set and classified patients smoking status
First step classification: smokers, non-smokers and unknowns
Second step classification: classified smokers into smokers with unknown intensity, light, intermediate, intermittent, past and heavy smokers
Test dataset: 34% (1,120 of 3,296) of randomly selected reference standard data
Tested trained machine learning classfiers on test dataset
Evaluated performance of machine learning classifiers by calculating precision, recall and f-1 measure
Applied machine learning classifier that performed best on the new dataset that is not manually annotated (49% (3,114 of 6,410))
Classified and reported number of smokers (light, intermediate, intermittent, past, unknown intensity, and heavy), non-smokers, unknowns
Confirmed SVM performance on new dataset by manually evaluating 315 patients' assigned smoking statuses

Fig. 1 Flowchart describing the sequence of steps involved in classifying patients' smoking status based on smoking history from one visit documented in the electronic dental record (EDR). SVM, support vector machine

extensible Human Oracle Suite of Tools for the manual annotation.[26,27] Next, we performed interannotator agreement to evaluate the agreement between two annotators. The agreement between the two annotators was 0.92 (Cohen's kappa),[28] which indicated an excellent agreement. They resolved any disagreements through discussion and consensus. The finalized annotations were considered as the reference standard.

Training Machine Learning Classifiers

In this study, we applied three ML classifiers such as naïve Bayes, SVM, and random forest, which were applied in previous studies to automatically extract information from free-text in the EHRs.[12,14–22,29–31] To apply classifiers, we utilized the Weka ML workbench[32] platform. We randomly split 66% (2,176 patient smoking histories) of the reference standard (annotated data) into a training set to train the classifiers and the remaining 34% (1,120) into a test set to evaluate these classifiers.

We applied the ML classifiers on the data sets in two steps. In the first step, the ML classifiers classified patients into three categories: (1) smokers, (2) nonsmokers, and (3) unknowns (see ►Fig. 1). In the second step, the ML classifiers classified patients in the smokers category based on their smoking intensity and duration: (1) smokers with unknown intensity, (2) past, (3) light, (4) intermediate, (5) intermittent, and (6) heavy smokers (see ►Fig. 1). This two-step approach provides a better result when compared with classifying all the smoking categories at once.[14]

We used the following functions, and features while applying the classifiers. We tokenized each annotated smoking history sentence using the string to vector function, which also converts the string into a set of attributes that contain word occurrence information. The NGram tokenizer with 1 to 2 window word feature produced single word (N1) and bigrams (N2) representations. The term frequency-inverse document frequency (TF-IDF) determined the word frequencies and the most important word/s within the

Table 1 Guidelines to manually annotate patient's smoking status and create reference standard

Smoker classification	Description and annotation guidelines	Examples of literal text matches from EDRs
Nonsmoker	If patient has never smoked tobacco	"Patient denies," "never," "N," "No tobacco"
Smokers with unknown intensity	If a patient was a smoker within the past year. Additionally, the history does not contain any detailed information such as the number of cigarettes or packs/day	"Patient smoke cigarettes"
Past smoker	If the patient was a smoker 1 year or more ago but who has not smoked for at least 1 year. Additionally, the history does not contain any detailed information such as the number of cigarettes or packs/day	"Patient smoked for 9 years and then quit 2 years back"
Light smoker	If the patient smokes less than or equal to 10 cigarettes/day	"Patient smokes 9 cigarettes a day," "Patient smokes 2 cigarettes a week"
Intermediate smoker	If the patient smokes in between 11 and 20 cigarettes per day	"Patient smokes 12 cigarettes per day"
Intermittent smoker	If the patient smokes occasionally or socially, such as in parties, bars, and nondaily basis	"Patient smokes occasionally," "Patient smokes socially," "smokes only in parties or holidays"
Heavy smoker	If the patient smokes more than or equal to 20 cigarettes or more than 1 pack per day	"25 cigarettes a day," "more than 1 pack a day"
Unknown	When there is no information present regarding patient's smoking status or blank space present within the history	"Patient drinks 5 beers a day"

Abbreviation: EDR, electronic dental record.

annotated training set that classifies a patient's smoking status. For example, in the sentence "patient smokes two cigarettes daily," "1 to 2" window word feature generated the following NGram tokens: (1) "patient smokes," (2) "smokes two," (3) "two cigarettes," and (4) "cigarettes daily," (5) "patient," (6) "smokes," (7) "two," (8) "cigarettes," and (9) "daily." The TF-IDF determined whether the word frequencies in a document should be transformed into: fij*log(num of Docs/num of Docs with word i), where "fij" is the frequency of word "i" in each sentence (instance), while j describes word occurrence in the training set. So, in the sentence, "patient smokes two cigarettes daily," the tokens, "smokes" and "two cigarettes" were ranked as the most important word features to classify a patient as "light smoker."

The classifiers used these word features to classify patients' detailed smoking status. ►Tables 2 and 3 list the top ranked word features utilized for each smoking status. The most frequent word features in the documents (see ►Tables 1 and 3) became identification term features for the classifiers to assign patients' smoking status. Finally, we utilized 10-fold cross-validation on the training data set, to obtain an unbiased model, and to mitigate overfitting the model.[15,20,21,29,30]

Evaluating ML Classifiers Performance on the Test Set and Applying it to the New Data Set

We evaluated the performance of the three ML classifiers by calculating their precision (true positive / (true positives + false positives)), recall (true positives / (true positive + false negative)), and F-measure (2 * (precision * recall) / (precision + recall)).[31] We selected the classifier that performed best to classify patients' smoking status from the new data set consisting of 3,114 unique patients' records. The purpose of applying trained classifier on the new data set was to extract patients' smoking information from a different data set using the trained ML classifier that demonstrated

Table 2 Word features utilized to classify patients into smokers, nonsmokers, and unknown categories

Nonsmokers	"no/No," "denies," and "N"
Smokers	"tobacco," "day," "smoking," and "socially"
Unknowns	"Alcohol," "alcohol," and "drinks"

Table 3 Word features utilized to classify smokers into light, intermediate, intermittent, heavy, smokers with unknown intensity, past, and intermittent categories

Light smoker	"1/2," "day," and "years"
Intermediate smokers	"pack," "1," and "years"
Heavy smokers	"a," "day," "pack," and "years"
Smokers with unknown intensity	"smoking," "years"
Past smokers	"quit," "ago"
Intermittent smoker	"socially," and "occasionally"

excellent performance. To confirm the classifier's performance on the new data set, we evaluated the performance on a subset of 315 randomly selected annotated patients' smoking histories.

Results

Based on the annotation guidelines (see ►Table 1), the reference standard consisted of 1,076 (33%) smokers, 1,300 nonsmokers (39%), and 920 (28%) unknowns (see ►Table 4). ►Table 4 also displays smokers classified based on their smoking intensity. ►Table 5 displays patients' smoking statuses in the new data set, and the smokers classified based on their smoking intensity. The new data set consisted of 1,090 (35%) smokers, 1,214 (39%) nonsmokers, and 810 (26%) unknowns in the new data set (see ►Table 5). Among smokers, most patients were classified as light smokers in both the reference standard (253 [24%]) and new data set (346 [31%]) (see ►Tables 2 and 3). Additionally, least number of patients were classified as heavy smokers in both reference standard (43 [4%]) and new data set (45 [4%]) (see ►Tables 1 and 3). Among the three ML classifiers, SVM performed excellent (precision, recall, and F-measure of 98%) (►Table 6) in classifying patients into smokers, nonsmokers, and unknowns. SVM also performed best in classifying smokers further based on their smoking intensity (precision: 88%, recall: 82%, F-measure: 84%) (see ►Table 7).

As demonstrated in ►Table 8, SVM achieved moderate performance (precision: 90%, recall: 44%, F-measure: 60%) in classifying heavy smokers, which could be due to the small sample size in our training set. SVM, in general, tends to be biased toward majority classes giving less priority to minority classes. Here, "heavy smoker" was considered minority class due to its small sample size in the data set. In contrast, SVM performed excellent in identifying intermittent smokers (precision: 95%, recall: 98%, F-measure: 96%) even though their sample size was low. The probable reason could be the consistency in the documentation of intermittent

Table 4 Number of smokers, nonsmokers, and unknowns in the reference standard

Smoking status and intensity		Reference standard (%)
Smokers	Light smoker	253 (24)
	Intermediate smoker	234 (22)
	Smoker with unknown intensity	164 (15)
	Past smoker	302 (28)
	Intermittent smoker	80 (7)
	Heavy smoker	43 (4)
Total smokers		1,076 (33)
Nonsmoker		1,300 (39)
Unknown		920 (28)
Total		3,296 (100)

Note: Percentages are rounded to the nearest integer.

Table 5 Number of smokers, nonsmokers, and unknowns in the new data set

Smoking status and intensity		New data set (%)
Smokers	Light smoker	346 (31)
	Intermediate smoker	314 (29)
	Past smoker	197 (18)
	Smoker with unknown intensity	129 (12)
	Intermittent smoker	59 (6)
	Heavy smoker	45 (4)
Total smokers		1,090 (35)
Nonsmoker		1,214 (39)
Unknown		810 (26)
Total		3,114 (100)

Note: Percentages are rounded to the nearest integer.

Table 6 Performance of machine learning classifiers on classifying patients into smoker, nonsmoker, and unknown categories

MLCs	Precision	Recall	F-measure
SVM	0.98	0.98	0.98
Random forest	0.96	0.96	0.96
Naïve Bayes	0.95	0.95	0.95

Abbreviations: MLCs, machine learning classifiers; SVM, support vector machine.

Table 7 Performance of machine learning classifiers to classify smokers into light smoker, intermediate smoker, intermittent smoker, heavy smoker, past smoker, smoker with unknown intensity, and heavy smoker

MLCs	Precision	Recall	F-measure
SVM	0.89	0.82	0.84
Random forest	0.75	0.73	0.73
Naïve Bayes	0.69	0.70	0.70

Abbreviations: MLCs, machine learning classifiers; SVM, support vector machine.

Table 8 Performance of support vector machine to classify patients based on their smoking intensity

Smoking status	Precision	Recall	F-measure
IntS	0.95	0.98	0.96
IS	0.90	0.88	0.89
PS	0.89	0.89	0.89
LS	0.87	0.87	0.87
SUI	0.76	0.86	0.81
HS	0.90	0.44	0.60
Average	0.89	0.82	0.84

Abbreviations: HS, heavy smoker; IntS, intermittent smoker; IS, intermediate smoker; LS, light smoker; PS, past smoker; SUI, smoker with unknown intensity.

Table 9 Number of patients’ correctly and incorrectly identified smoking intensity by the support vector machine in the reference standard

Patients’ smoking statuses identified by SVM	Patients’ smoking statuses identified by manual annotation								Total smokers
	Classification	SUI	HS	IS	IntS	LS	PS	Total	
	PS	5	0	0	0	0	**297**	302	1,076
	LS	15	0	14	2	**221**	1	253	
	IS	13	0	**207**	1	11	2	234	
	SUI	**141**	2	3	0	5	13	164	
	IntS	8	0	1	**67**	4	0	80	
	HS	3	**19**	7	1	12	1	43	

Abbreviations: HS, heavy smoker; IntS, intermittent smoker; IS, intermediate smoker; LS, light smoker; PS, past smoker; SUI, smoker with unknown intensity; SVM, support vector machine.
Note: Bold numbers indicate correctly identified smoking status.

smoking status in the EDR. As described in our annotation guidelines (see ►Table 1), we classified patients into this category when they smoked occasionally in parties and clubs. Many of these histories contain words such as “occasionally,” “socially,” and “in party.” Frequent appearance of these words helped the SVM to differentiate this class from others.

►Table 9 describes correctly and incorrectly identified smoking statuses by SVM in our test set. At times, the classifier could not differentiate between smokers with unknown intensity and past smoker, although smokers with unknown intensity status was written in present tense (“smokes,” “smoke”), and past smoker in past tense (“smoked”). However, SVM differentiated sentences describing “nonsmokers” with negation attributes such as “patient never smokes,” or “patient does not smoke” from those describing smokers with affirmed attributes such as “patient smokes,” or “patient is smoking since last 20 years.”

In the new data set, SVM achieved precision of 0.96, recall of 0.96, and F-measure of 0.96 to classify patients into smokers, nonsmokers, and unknowns. SVM achieved precision of 0.90, recall of 0.84, and F-measure of 0.88 classifying patients into light, intermediate, heavy, intermittent, past, and smokers with unknown intensity category.

Discussion

The study results demonstrated the feasibility of automatically classifying patients’ smoking statuses including smoking intensity from their EDR that may not be easily available from the EHR data. The ML classifiers achieved excellent performance to classify patients into smokers, nonsmokers, or unknowns because of a large training set.[33] Also, the consistency in the documentation of smoking status across EDRs enabled ML to recognize patterns of smoking documentation and led to superior performance. Compared with previous studies in medicine,[16,18,21] the SVM performed superiorly in classifying smokers versus nonsmokers. While previous studies[15,16,18,21] achieved F-measure, precision, and recall of less than 96% (except one that achieved a precision of 98%), our classifier achieved a higher F-measure, precision, and recall of 98%. The classifier also distinguished records, which did not have any smoking information into “unknowns.”

We found no major difference between SVM’s performance on the test set and new data set confirming its ability to classify patients’ smoking statuses based on smoking intensity. We observed that SVM correctly classified patients’ smoking status when the most important and frequent word features (see ►Tables 2 and 3) were present in the sentence. However, it incorrectly classified when it did not recognize a writing pattern observed rarely in the training set such as “x packs per month” instead of “x packs per week or day.”

A significant strength of this study is leveraging the presence of a separate section for smoking in the EDR, which eliminated preprocessing of clinical notes significantly. Previous work on retrieving smoking status first retrieved clinical notes that described a patient’s smoking status. Next, they extracted sentences, which defined smoking status from the clinical notes, and in the last step, classified patients’ smoking status based on the context found in the sentence. Due to this need for extensive preprocessing and multiple steps, systems demonstrated low to moderate performance in classifying patient’s smoking status.[14,15,17,18] We also did not require extracting smoking-specific sentences from lengthy clinical notes, which eliminated the first two steps and thus enhanced system performance with fewer errors.

Unlike previous studies[13–17,21] which focused mainly on classifying patients as past, current, and nonsmokers, we also classified patients based on their smoking intensity. To date, only one study by Wang et al classified patients’ smoking status based on smoking intensity by adding two additional categories such as light and heavy smokers.[18] In this study, we took one step further and added intermediate smoker categories because recent studies indicate a lower prevalence of heavy smokers and a higher prevalence of intermediate smokers. The study results confirmed this finding with only 4% heavy smokers and maximum light smokers followed by intermediate smokers (see ►Tables 1 and 3). Retrieving detailed smoking status is beneficial not only for research purposes but also for clinical care purposes. Awareness of patients’ smoking status enables clinicians to determine their tobacco dependency levels and interest in receiving counseling to quit smoking based on their smoking intensity.[34,35] They are also able to determine their patients' risk of developing dental diseases and prognosis following dental treatments especially surgical procedures.[36,37]

The study results indicated the potential of leveraging EDR data to obtain detailed smoking history such as smoking intensity that may not be present in the EHR. In addition, detailed smoking history is not easily retrievable from the EHR.[15,19] Medical providers could benefit from the detailed smoking history present in the EDRs because smoking is also an established risk factor for many systemic diseases such as lung cancer and cardiovascular diseases. Recently, there is growing awareness and research on the value of integrating dental and medical records to coordinate care and to investigate the potential association between oral and systemic diseases.[38] As this initiative progresses, it is worthwhile to determine the value of sharing social habits such as smoking and diet to increase dental and medical providers' awareness of their patients' major risk factors for common chronic diseases.

A limitation of these ML classifiers is that it may not perform well on other institutional data because they were trained on our institutional data. The extent of smoking documentation and writing patterns may also vary in different clinic settings. However, the approach utilized in this study to train ML classifiers could be extended to data from other settings, and thus researchers do not have to start from the beginning.

Conclusion and Future Work

This study demonstrated the feasibility of extracting patients' detailed smoking status automatically from EDR. EDR data could serve as a valuable source for obtaining patients' detailed smoking information based on smoking intensity. Although our ML classifier performed excellently in classifying patients into light, intermediate, smokers with unknown intensity, intermittent, nonsmoker, and unknown, we need to enhance their performance to classify heavy smokers. We will enhance the classifier's performance by utilizing methods such as oversampling and Synthetic Minority Over-sampling Technique. We also plan to test this classifier on data from other dental and medical settings. Additionally, we plan to study changes in patients' smoking status from their longitudinal smoking records. Finally, we will use this classifier to report the prevalence of smoking in our patient population and to determine the correlation of smoking with dental diseases such as dental caries and periodontal disease.

Human Subjects Protections

This study was reviewed and approved by our institutional review board (IRB: 1310579094) and granted authorization.

Funding

This study was funded by Dr. Thankam Paul Thyvalikakath's start-up funds from Indiana University School of Dentistry.

Conflict of Interest

None.

Acknowledgments

The authors would like to thank Mr. Craig Eberhardt for generating study data set and Heng Yi Wu for providing Weka training.

References

1 Chatzopoulos G. Smoking, smokeless tobacco, and alcohol consumption as contributing factors to periodontal disease. Northwest Dent 2016;95(01):37–41
2 Marcenes W, Kassebaum NJ, Bernabé E, et al. Global burden of oral conditions in 1990-2010: a systematic analysis. J Dent Res 2013; 92(07):592–597
3 Durham J, Fraser HM, McCracken GI, Stone KM, John MT, Preshaw PM. Impact of periodontitis on oral health-related quality of life. J Dent 2013;41(04):370–376
4 Martinez-Canut P, Lorca A, Magán R. Smoking and periodontal disease severity. J Clin Periodontol 1995;22(10):743–749
5 Kinane DF, Chestnutt IG. Smoking and periodontal disease. Crit Rev Oral Biol Med 2000;11(03):356–365
6 Morse DE, Psoter WJ, Cleveland D, et al. Smoking and drinking in relation to oral cancer and oral epithelial dysplasia. Cancer Causes Control 2007;18(09):919–929
7 Charangowda BK. Dental records: an overview. J Forensic Dent Sci 2010;2(01):5–10
8 Chaffee BW, Couch ET, Ryder MI. The tobacco-using periodontal patient: role of the dental practitioner in tobacco cessation and periodontal disease management. Periodontol 2000 2016;71(01): 52–64
9 Rush WA, Schleyer TK, Kirshner M, et al. Integrating tobacco dependence counseling into electronic dental records: a multimethod approach. J Dent Educ 2014;78(01):31–39
10 Song M, Liu K, Abromitis R, Schleyer TL. Reusing electronic patient data for dental clinical research: a review of current status. J Dent 2013;41(12):1148–1163
11 Siddiqui Z, Wang Y, Makkad P, Thyvalikakath T. Characterizing restorative dental treatments of Sjögren's syndrome patients using electronic dental records data. Stud Health Technol Inform 2017;245:1166–1169
12 Wu Y, Rosenbloom ST, Denny JC, et al. Detecting abbreviations in discharge summaries using machine learning methods. AMIA Annu Symp Proc 2011;2011:1541–1549
13 Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc 2008;15(01):36–39
14 Cohen AM. Five-way smoking status classification using text hotspot identification and error-correcting output codes. J Am Med Inform Assoc 2008;15(01):32–35
15 Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. Conf Proc IEEE Eng Med Biol Soc 2014;2014:2710–2713
16 Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform 2015;58(Suppl):S67–S77
17 Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(01):14–24
18 Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer Inform 2016;15:237–242
19 De Silva L, Ginter T, Forbush T, et al. , eds. Extraction and quantification of pack-years and classification of smoker information in semi-structured Medical Records. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA; 2011

20 Sohn S, Savova GK. Mayo clinic smoking status classification system: extensions and improvements. AMIA Annu Symp Proc 2009;2009:619–623

21 Jonnagaddala J, Dai H-J, Ray P, Liaw S-T. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In: Proceedings of the BioNLP 15; 2015:147–151

22 Schane RE, Ling PM, Glantz SA. Health effects of light and intermittent smoking: a review. Circulation 2010;121(13):1518–1522

23 Schoenborn CA, Adams PE. Health behaviors of adults: United States, 2005-2007. Vital Health Stat 10 2010;(245):1–132

24 Neumann T, Rasmussen M, Heitmann BL, Tønnesen H. Gold standard program for heavy smokers in a real-life setting. Int J Environ Res Public Health 2013;10(09):4186–4199

25 Neves M, Leser U. A survey on annotation tools for the biomedical literature. Brief Bioinform 2014;15(02):327–340

26 South BR, Mowery D, Suo Y, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014;50:162–172

27 South BR, Shen S, Leng J, Forbush TB, DuVall SL, Chapman WW, eds. A prototype tool set to support machine-assisted annotation. In: Proceedings of the 2012 Workshop Biomed Natural Language Processing: Association for Computational Linguistics; 2012

28 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282

29 Carroll RJ, Eyler AE, Denny JC. Naïve electronic health record phenotype identification for rheumatoid arthritis. AMIA Annu Symp Proc 2011;2011:189–196

30 Castro VM, Minnier J, Murphy SN, et al; International Cohort Collection for Bipolar Disorder Consortium. Validation of electronic health record phenotyping of bipolar disorder cases and controls. Am J Psychiatry 2015;172(04):363–372

31 Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(03):296–298

32 Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD. 2009;11(01):10–18

33 Wei Z, Wang W, Bradfield J, et al; International IBD Genetics Consortium. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet 2013;92(06):1008–1012

34 Clinical Practice Guideline Treating Tobacco Use and Dependence 2008 Update Panel, Liaisons, and Staff. A clinical practice guideline for treating tobacco use and dependence: 2008 update. A U.S. Public Health Service report. Am J Prev Med 2008;35(02):158–176

35 Chambrone L, Preshaw PM, Rosa EF, et al. Effects of smoking cessation on the outcomes of non-surgical periodontal therapy: a systematic review and individual patient data meta-analysis. J Clin Periodontol 2013;40(06):607–615

36 Baig MR, Rajan M. Effects of smoking on the outcome of implant treatment: a literature review. Indian J Dent Res 2007;18(04):190–195

37 Kotsakis GA, Javed F, Hinrichs JE, Karoussis IK, Romanos GE. Impact of cigarette smoking on clinical outcomes of periodontal flap surgical procedures: a systematic review and meta-analysis. J Periodontol 2015;86(02):254–263

38 Atchison KA, Weintraub JA, Rozier RG. Bridging the dental-medical divide: case studies integrating oral health care and primary health care. J Am Dent Assoc 2018;149(10):850–858