UTILIZING ELECTRONIC DENTAL RECORD DATA TO TRACK PERIODONTAL DISEASE CHANGE

Jay Sureshbhai Patel

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

July 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Josette Jones, RN, PhD, Chair

_____
Thankam Paul Thyvalikakath, DMD, MDS, PhD

January 15, 2020

_____
Mathew Palakal, PhD

_____
Michael Kowolik, BDS, PhD

_____
Radha Nagarajan, PhD

Dedication

To my mother Varsha Patel, father Suresh Patel, sister Nisha Patel, brother in law Charchil Vejani, nieces Prashi Vejani and Pearl Vejani, and Brian McDevitt.

Acknowledgment

The author would like to acknowledge first and foremost his dissertation committee members Dr. Thankam Paul Thyvalikakath, Dr. Josette Jones, Dr. Mathew Palakal, Dr. Michael Kowolik, and Dr. Radha Nagarajan for all of their guidance, assistance, and support. The author specially wants to thank his research advisor Dr. Thankam Paul Thyvalikakath, for her mentorship, and teaching him important skills such as critical thinking and scientific writing. The author also would like to special thank Dr. Josette Jones for being his academic advisor and mentorship. The author shows appreciation to Indiana University School of Dentistry for supporting him through Dr. Thyvalikakath's start-up funds, and School of Informatics and computing for tuition scholarship throughout his Ph.D. program. The author also wishes to acknowledge Mr. Ahad Zai, Mr. Krishna Kumar, and Mr. Craig Eberhardt for their help with the development of the computer algorithms. Next, the author would like to show his appreciation to Dr. Daniel Shin and Dr. Lisa Willis for manually reviewing patient records to evaluate the performance of the computer algorithms. The author wants to also thank Dr. Anand Kulanthaivel, Dr. Shuning Li, Dr. Michelle Lapradd, Lukas Miller who provided him their guidance during the qualifying exam, proposal defense, dissertation writing and defense. The author wants to thank Mr. Sean Stone to help with educating the literature search resources available at Indiana University library. The author thanks Ms. Elizabeth Cassell, Ms. Robyn Hart, and Ms. Debra Barker for their administration support and help. Finally, the author wishes to acknowledge the phenomenal amount of unconditional moral support given by his friends, family and mentors throughout the past five years in achieving the difficult tasks of completing this Doctorate degree.

Jay Sureshbhai Patel

UTILIZING ELECTRONIC DENTAL RECORD DATA TO TRACK PERIODONTAL DISEASE CHANGE

Periodontal disease (PD) affects 42% of US population resulting in compromised quality of life, the potential for tooth loss and influence on overall health. Despite significant understanding of PD etiology, limited longitudinal studies have investigated PD change in response to various treatments. A major barrier is the difficulty of conducting randomized controlled trials with adequate numbers of patients over a longer time. Electronic dental record (EDR) data offer the opportunity to study outcomes following various periodontal treatments. However, using EDR data for research has challenges including quality and missing data. In this dissertation, I studied a cohort of patients with PD from EDR to monitor their disease status over time. I studied retrospectively 28,908 patients who received comprehensive oral evaluation at the Indiana University School of Dentistry between January 1st-2009 and December 31st-2014. Using natural language processing and automated approaches, we 1) determined PD diagnoses from periodontal charting based on case definitions for surveillance studies, 2) extracted clinician-recorded diagnoses from clinical notes, 3) determined the number of patients with disease improvement or progression over time from EDR data. We found 100% completeness for age, sex; 72% for race; 80% for periodontal charting findings; and 47% for clinician-recorded diagnoses. The number of visits ranged from 1-14 with an average of two visits. From diagnoses obtained from findings, 37% of patients had gingivitis, 55% had moderate periodontitis, and 28% had severe periodontitis. In clinician-recorded diagnoses, 50% patients had gingivitis, 18% had mild, 14% had moderate, and 4% had severe periodontitis.

The concordance between periodontal charting-generated and clinician-recorded diagnoses was 47%. The results indicate that case definitions for PD are underestimating gingivitis and overestimating the prevalence of periodontitis. Expert review of findings identified clinicians relying on visual assessment and radiographic findings in addition to the case definition criteria to document PD diagnosis.

Josette Jones, RN, PhD, Chair

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAP: American Academy of Periodontology

CDC: Center for Disease Control

COE: Comprehensive Oral Evaluation

EDR: Electronic Dental Record

EHR: Electronic Health Record

GI: Gingival Index Score

IUSD: Indiana University School of Dentistry

NHANES: National Health and Nutrition Examination Survey

PD: Periodontal Disease

PF: Periodontal Form

PM: Periodontal Maintenance

POE: Periodic Oral Evaluation

PPD: Periodontal Pocket Depth

BOP: Bleeding on Probing

CAL: Clinical Attachment Loss

PRE: Periodontal Re-evaluation

1: Introduction

Despite advances in periodontal disease (PD) research and periodontal treatments, 42% of the U.S. population has PD, which can lead to tooth loss, poor quality of life, and increased healthcare cost [33]. To date, limited studies show the effectiveness of current periodontal treatments in preventing the progression of the disease and tooth loss based on patient characteristics [53, 81, 82]. A major barrier is the difficulty of conducting randomized controlled trials with adequate numbers of patients over a longer time because of several reasons such as ethical reasons, expenses, and difficulty in enrollment and retaining patients for a longer time [25]. Moreover, it is also well studied that the PD can be prevented if the risk factors responsible for PD progression could be controlled by assessing patients' disease risk [57, 87, 88, 89, 123]. As a result, prediction models to assess patient specific disease risk have been developed [57, 58, 68, 78, 87, 89, 96]. However, studies [49, 68, 102, 119] have shown that these tools are not representing the current patient population and unable to provide patient specific disease risk and treatment recommendations.

The increased availability of longitudinal patient care data electronically through the electronic dental record (EDR) offers an opportunity to characterize present patient population's demographics, disease profiles, periodontal treatment outcomes, and develop prediction models with up-to-date information [25, 110, 115]. However, EDRs are designed to support patient care and not research purposes, therefore, some challenges exist while using their data for research. They include 1) questionable quality of data [120], 2) specific disease information documented in multiple sections of the EDR (e.g., disease diagnosis could be either recorded using diagnostic codes and/or in progress notes), and 3) missing information [16, 23, 25, 36, 47, 133] compared to prospective study which typically has designed to

study associations, causations, and treatment outcomes, EDRs may or may not be able to provide as granular data as in prospective study. For instance, a prospective study [33] by the National Health and Nutrition Examination Survey (NHANES) examining the nationwide prevalence of PD in the United States captured each patient's PD clinical variables such as patient demographics, clinical attachment loss, and probing depth information by examining four sites per tooth. In contrast, in the EDR, the capture of these relevant clinical variables can be sporadic and varied in quality and detail. A patient's PD status could be recorded in the progress notes or/and in the diagnosis field using diagnostic codes. At the same time, utilizing Electronic Health Record (EHR) data has several advantages over data collected through prospective studies such as EDR that could provide patient's longer follow-up data that is difficult, and expensive to study through prospective study design [25].

Therefore, before using EHR and EDR data for research, it is important to first determine the data quality measures such as completeness and concordance of clinical variables because if the data used to study a research question is flawed then the output will be flawed as well [47]. At the same time, it is also important to utilize all sections of the EDR to find patient's relevant information as information could be reported in multiple sections of the electronic health record. [37, 42, 63].

Therefore, the objective of this dissertation was to, first, evaluate the quality of the EDR data and its suitability for PD research and, second, to determine the feasibility of using longitudinal EDR data to monitor PD change over time. Two data quality measures, completeness, and concordance were used to evaluate the EDR data quality. A record was considered complete if all findings that contribute to PD diagnoses such as clinical attachment loss, bleeding on probing, periodontal pocket depth, and PD diagnosis were

recorded in the EDR. Concordance was measured by determining the agreement between diagnosis recorded by clinicians and diagnoses derived from periodontal charting findings in the EDR. These objectives are achieved by conducting the following aims.

Aim 1: Determine the completeness and concordance of patients' periodontal disease diagnosis recorded during comprehensive oral evaluation in their electronic dental record.

The hypothesis of this aim is that the completeness of PD diagnosis information reported in the EDR data would be at least 80% complete. The value of completeness is set of 80% because at the Indiana University School of Dentistry (IUSD), dental students are mandated to record periodontal findings derived from periodontal charting. It is also hypothesized that the concordance between PD diagnoses generated from periodontal charting and clinician-documented diagnoses would be at least 90%. The value of 90% was set for concordance because all patients are required to have complete PD diagnosis and findings documentation while receiving a comprehensive oral evaluation in the EDR. Moreover, periodontal findings utilized to determine PD diagnosis from charting data were same to the periodontal findings used by the clinicians at IUSD while documenting PD diagnosis.

To achieve this aim, the following steps were performed. First, a patient cohort was generated from the IUSD clinics. Second, patients' PD diagnoses were generated from the periodontal charting findings (bleeding on probing, clinical attachment loss, and periodontal pocket depth). Next, clinician-recorded PD diagnoses were retrieved from the "diagnosis" section of the periodontal evaluation form. Next, the completeness of EDR variables including patient demographics, periodontal charting, and PD diagnosis in the periodontal evaluation form was calculated. Lastly, the concordance was determined

between PD diagnoses derived from periodontal charting and clinician recorded diagnoses in the periodontal evaluation form. The results of this study would contribute to determining the quality of periodontal information recorded for patients who underwent COE in the EDR data. This study results would also contribute to determining the appropriateness of utilizing EDR data for research to assess treatment outcomes and to predict PD.

Aim 2: Determine the feasibility of tracking changes in a patient's periodontal disease diagnosis using electronic dental record data.

It is hypothesized that EDR data would provide PD diagnosis for at least 50% patients who underwent COE to track their disease over time. In order to determine the PD change, at least two visits must be present in the EDR data.

To test this hypothesis, first, average visits for patients who received COE between January 1, 2009 to December 31, 2014 were generated. Next, a computer algorithm was developed that generated the number of patients whose disease status did not change over time, and patients whose disease status regressed or progressed over time. This study results will help in determining the potential of EDR data to predict patient's disease progression over time and determine PD treatment outcomes.

1.1: Significance

Since the last two decades, there is a huge shift among informatics researchers to use EHR data for clinical research. Since the EHR data has not been collected for research purposes, and for patient care purposes, researchers have worked on developing advanced machine learning algorithms, artificial intelligence, and statistical models to utilize EHR data to extract information, predict disease risk, and to provide personalized treatment recommendations. Despite this huge shift, the transition of the research results generated through EHR data to practice is limited and controversial. This is because there are many challenges associated with the EHR data such as questionable quality, missing information, and questionable reliability [133]. The first law of informatics is *"the data should only be used for which it has been collected for* [127]*",* as, EHR data has not been collected for research purpose, the quality matters and it must be tested before its intended use. It is necessary to understand and evaluate the quality of the data before developing prediction models using machine learning algorithms and artificial intelligence to avoid errors, biases, and flawed outcomes inherent the data. A recently published (05 June 2020) news *"High-profile coronavirus retractions raise concerns about data oversight* [83]*"* from nature sets good example of the importance of evaluating the quality of the data before its use.

In this dissertation, informatics methods have been developed to evaluate the quality of the EDR data before its intended use for research. The informatics approaches developed in the study would determine the completeness of PD diagnosis information available in the EDR data to study periodontal treatment outcomes and predict periodontal disease risk and prognosis. Moreover, the results of this study would determine the feasibility of monitoring change in patients' PD diagnosis over time. Informatics

algorithms developed in this dissertation can automatically determine patients' PD diagnosis from periodontal findings and also extract information automatically from clinician-recorded diagnoses. Without the informatics methods, it would be humanly impossible to process big EHR data that otherwise would require manual review that is time consuming, expensive, labor intensive, and error prone. Researchers from other institutes will be able to use informatics algorithms developed in this dissertation to evaluate the quality of the EDR data and to generate diagnosis from their patients' periodontal charting findings. This would allow them to use EDR data for clinical research and quality assurance purposes. Therefore, this dissertation sets a model and a process to evaluate the quality of the EDR data using advanced informatics approaches.

Another significant outcome of this dissertation is that, it demonstrated the comprehensive steps needed to examine the quality of the EHR data. Studies [36, 133] demonstrating EHR data quality are theorical models, however, the process and frame work of utilizing theoretical model to practice was demonstrated in this dissertation. Researchers from different healthcare fields such as medicine, physical therapy, nutritional science, pharmaceutical science, and even other non-healthcare related fields will be able to utilize this framework to evaluate the quality of their electronic data before its intended use. As a result, electronic data for research can be utilized optimally, and bias generated through flawed data can be minimized through conducting regular data quality assurance through the framework proposed in this dissertation.

2: Background

2.1 Periodontal Disease Etiology and Prevalence

The life expectancy of people in the United States has been increasing when compared to three decades ago [79, 130]. People are living longer and retaining more natural teeth because of the advancement of dental research and treatments [30]. However, the older population is often suffering from chronic systemic diseases such as cardiovascular disease, diabetes, osteoporosis and dental diseases such as periodontal disease (PD) [39, 40].

PD is the seventh most common chronic oral condition affecting the gum tissue and bones supporting the teeth [33, 41]. According to the 2018 nationwide epidemiological study [33] examining the prevalence of PD reported 42% of the US adults suffering from PD . Among 42% of patients, approximately 8% of patients suffer from severe PD [33]. Compared to PD prevalence from two decades ago, the prevalence of PD still remains high among US adults. However, because of the availability of advanced treatment and frequent follow-ups, the dental community successfully able to move patient population from severe periodontitis to moderate periodontitis cases [29].

PD leads to the destruction of connective tissue and bone surrounding the tooth and can lead to other serious consequences, such as tooth loss. Research has shown that tooth loss is significantly associated with poor quality of life, especially in the older population [50]. Therefore, it negatively affects psychosocial behaviors, functionality limitations, physical pain and disability, discomfort, and psychological disability [43, 50]. In addition, it interferes with eating, speaking, cleaning, sleeping, smiling, working, and enjoying social contact [34, 43].

The etiology of PD involves the triad of bacterial infection, host inflammation, and risk factors. The bacteria associated with PD are predominantly gram-negative anaerobic bacteria and may include A. actinomycetemcomitans, P. gingivalis, P. intermedia, B. forsythus, C. rectus, E. nodatum, P. micros, S. intermedius, or Treponema sp. These bacteria are responsible for generating host inflammation, which leads to bone loss [17, 95]. The risk factors for PD are classified as causal, intermediate, and predisposing factors [4]. Causal risk factors have a direct effect on the likelihood of clinical events. Diabetes, smoking, and periodontal pathogens are causal risk factors for PD. Intermediate risk factors (also called risk markers, risk indicators, or putative risk factors) influence PD initiation and progression by having a synergetic effect on other risk factors, but not having a direct effect. Intermediate risk factors include immune dysfunction, poor oral hygiene behaviors, obesity, poor diet, stress, osteoporosis, and cognitive disorders. Predisposing factors have remote and rather complex associations with PD . The distal risk factors include age, gender, physical activity, social income status, education level, and social and physical environment. For example, lower educational level is closely correlated to lesser access to dental care services, a lower degree of periodontal health awareness, and irregular or deficient oral self-care practices, all of which are linked to poor oral hygiene habits that may lead to higher levels of dental plaque [4, 13, 128]. Risk factors can also be classified into modifiable and nonmodifiable risk factors. Nonmodifiable risk factors include age, gender, genetics, and ethnicity. Modifiable risk factors can be modifiable such as lifestyle behavioral factors (smoking, alcohol) and systemic diseases (diabetes) [4, 13, 39, 128]. Clinicians are mostly interested in modifiable risk factors and their prevention can reduce PD initiation and progression [13, 40].

In summary, because of extensive research on PD treatments and increased awareness of regular mechanical removal of plaque, patients with severe periodontitis and tooth loss have been reduced significantly compared to three decades ago. However, despite the advanced understanding of the disease, PD prevalence still remains high, especially cases with moderate severity. This chapter is important for this dissertation because it gives a summary of PD etiology, prevalence, and risk factors such as diabetes, cardiovascular diseases, and smoking that are responsible for influencing PD initiation and progression. It highlights the importance of patient education to control risk factors because the removal of bacterial infection solely is not enough to control the disease. Moreover, these factors may also have a negative influence on the PD treatment outcome. This chapter highlights the overall problem regarding the existing knowledge about PD and the importance of studying the disease that is attempted through this dissertation research.

2.2 Periodontal Disease Diagnosis and Classification

Classification and diagnosis are critical components in treating patients because they provide a framework to study the etiology, pathogenesis, prognosis, and treatment of a disease [8, 10]. In addition, such systems give clinicians a method of organizing the health care needs of their patients [6].

In this dissertation, a cohort of patients with PD was generated by determining PD diagnoses from periodontal findings and clinician-recorded diagnoses. Therefore, it is important to understand how the PD classification have evolved over the past two decades specially about the disease severity categories. In the following section, first, the evolution of PD classification over the past three decades is explained. The PD classification systems used in the past and current in use classification system is also described in this section. These classifications are originally represented in the studies [9, 10, 124].

In 1989, the first-time scientists and clinicians in the field of periodontology agreed upon a classification system for PD was at the World Workshop in Clinical Periodontics. Subsequently, a simpler classification was agreed upon at the 1st European Workshop in Periodontology [136]. The 1989 PD classification included different stages of periodontitis such as early-onset periodontitis, pre-pubertal periodontitis (localized, generalized), juvenile periodontitis (localized, generalized), rapidly progressive periodontitis, adult periodontitis, necrotizing ulcerative periodontitis, and refractory periodontitis. Unfortunately, the 1989 classification had many shortcomings including: 1) considerable overlap in disease categories, 2) absence of a gingival disease component, 3) inappropriate emphasis on the age of onset of disease and rates of progression, and 4) inadequate or unclear classification criteria. On the other hand, the 1993 European classification lacked

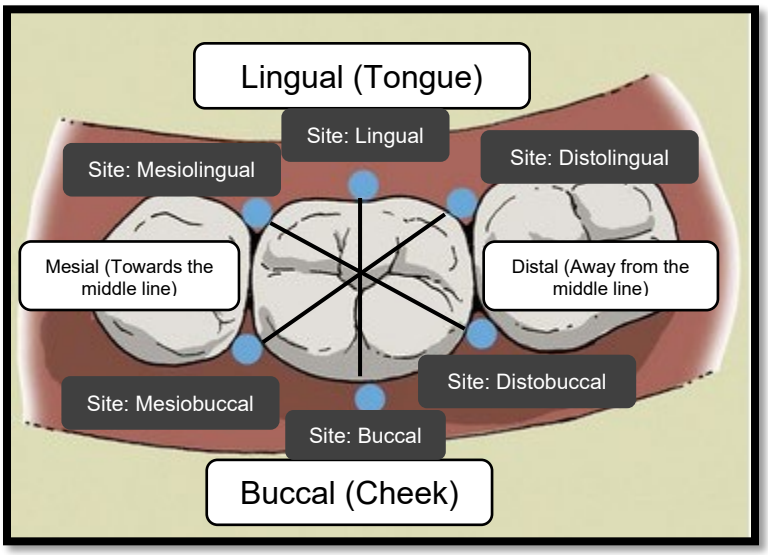
the detail necessary for adequate characterization of the broad spectrum of PD cases that are encountered in the clinical practice [8, 9]. To overcome these drawbacks, in 1999, the classification was revised significantly [10]. The major changes in the classification include the addition of the gingival disease component.

2.2.1 Gingivitis classification

Gingivitis is a mild form of PD, which is caused by the accumulation of plaque and calculus (tartar) surrounding the tooth. Gingivitis can be represented as swollen, red, and frequently bleeding gums [126]. Gingivitis is usually reversible if good oral hygiene is maintained, including regular dental visits, brushing, and flossing [8]. Gingivitis is typically not associated with the loss of connective tissue. When gingivitis is left untreated, it may lead to periodontitis, which is the more serious form of PD. One important feature of the section on dental plaque-induced diseases is gingivitis can be modified by 1) systemic factors such as perturbations in the endocrine system, 2) medications, and 3) malnutrition [9, 10].

To diagnose gingivitis, two methods have been most prevalently used, 1) gingival index score (GI), and 2) bleeding on probing score (BOP score) [126]. The GI score is measured using the visual assessment of gingival characteristics such as edema/swelling, redness, and the tendency of the marginal gingiva to bleed upon mechanical stimulation by a periodontal probe. Based on these findings, GI scores are generated on a 4-point ordinal scale: 0 = absence of inflammation; 1 = mild inflammation – slight change in color and little change in texture; 2 = moderate inflammation – moderate glazing, redness, edema, and hypertrophy; bleeding on pressure; 3 = severe inflammation – marked redness and hypertrophy, ulceration with tendency to spontaneous bleeding. GI is typically recorded based on four areas which include buccal, lingual, mesial and distal for each of six index teeth (maxillary right first molar and lateral incisor, maxillary left first premolar, mandibular left first molar and lateral incisor, mandibular right first premolar –called "Ramfjord teeth") [126].

The second method used to diagnose gingivitis is by calculating the BOP score. A BOP score is assessed as the proportion of bleeding sites (dichotomous yes/no evaluation) when stimulated by a standardized (dimensions and shape) manual probe with a controlled (~25 g) force to the bottom of the sulcus/pocket at six sites (mesiobuccal, buccal, distobuccal, mesiolingual, lingual, distolingual) on all present teeth. Illustration of the six periodontal probing sites is demonstrated in Figure 1. The BOP score is used for classifying patients into healthy, localized and generalized gingivitis. Recently AAP created a case definition and diagnostic consideration of plaque-induced gingivitis [126]. This review suggested that the BOP score has several advantages over GI score that include 1) GI score is subjective and visual assessment varies by clinicians' expertise, while, BOP assessment is objective, universally accepted, reliable and accurate clinical sign. 2) GI score is time-consuming for record-keeping, while, BOP score is easily assessed and recorded. Therefore, in this dissertation, the BOP score was utilized to diagnose patients' gingivitis status as described in the study [126].



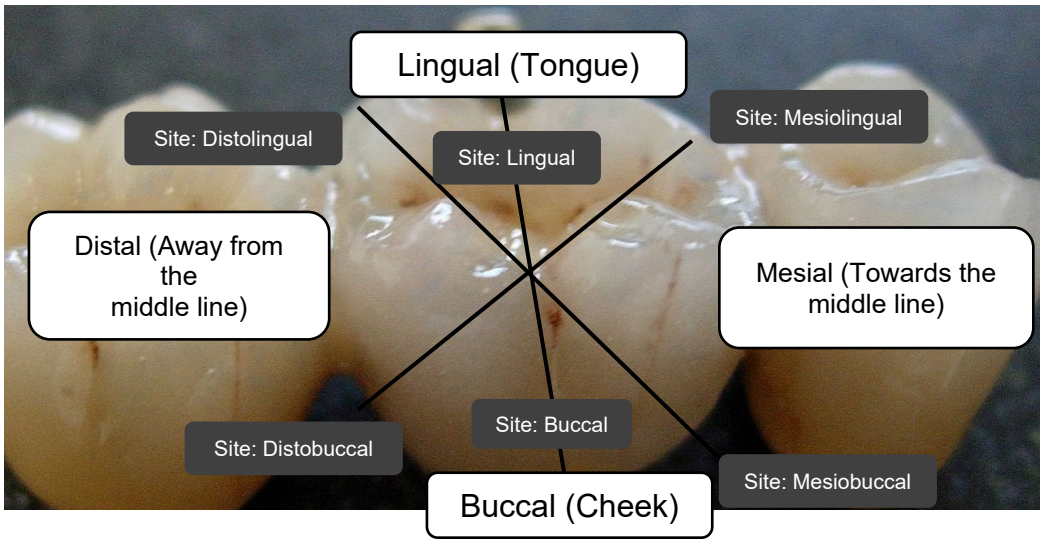


Figure 1: Illustration of six periodontal probing sites, mesiolingual, lingual, distolingual, distobuccal, buccal, mesiobuccal

(Citation: https://commons.wikimedia.org/wiki/File:Bridge_from_dental_porcelain.jpg, Six_Pocket_Sites.jpg, on (12/21/2019). Used under a Creative Commons 3.0 License.)

2.2.2 Periodontitis classification

Periodontitis is a severe form of PD in which there is an inflammation of the tissue surrounding the tooth [39]. Periodontitis is irreversible and if left untreated, it leads to tooth loss [18]. The classification of periodontitis has been changed frequently from 1982 [8, 9, 10] to 2018 [124]. In the following section, first the 1999 classification [10] is described which was used to diagnose periodontitis cases until the year 2017 at IUSD. Then, the updated periodontitis classification is described that was published in the year 2018 [124]. In the 1999 classification [10], periodontitis was classified based on its etiology and severity (see Table 1). In terms of etiology, periodontitis is classified as plaque-induced periodontitis and non-plaque-induced periodontitis. Periodontitis is also sub-classified into (1) chronic periodontitis, (2) aggressive periodontitis, (3) necrotizing periodontitis, (4) abscess of the periodontium, and (5) periodontitis associated with endocrine disorders. The severity of periodontitis is classified into mild, moderate, and severe forms, based on the amount of clinical attachment loss (e.g., the gingival margin migrates toward the root surface) and disease severity level. In mild periodontitis, the pockets deepen and there is more destruction of bone. Some clinical signs and symptoms include bad breath, bleeding while probing, and having a probing depth of 4 to 5 mm (usually measured with an instrument). In moderate periodontitis, the bad breath worsens and probing depth increases to 6 to 7 mm, which is a sign of more bone loss. At this stage, the type of accumulating bacteria changes and the bacteria enters the bloodstream, stressing the immune system. If moderate periodontitis is left untreated, it transforms into severe periodontitis, and probing depth increases to 7+ mm. This classification included the assessment of 1) periodontal charting findings (clinical attachment loss (CAL), bleeding on probing (BOP), periodontal pocket

depth (PPD), 2) radiographic bone loss information including severity and pattern of bone loss, clinical signs of inflammation, location and severity of plaque and calculus, and other relevant signs and symptoms such as pain or ulceration [10, 135]. The AAP task force report on the update to the 1999 classification of PD is described in Table 1 [6]. At IUSD, clinicians have used this classification until the year 2017 to diagnose patients' PD status.

A recently published study in 2018 [124] on redefining the classification of PD discovered some drawbacks of the 1999 classification. These drawbacks include, limited evidence on the differences in pathophysiology between chronic and aggressive periodontitis. There is evidence of multiple factors such as risk factors, bacterial infection and host inflammation that influence clinically observable disease outcomes at the individual level. This is true for both chronic and aggressive periodontitis cases. Second, a classification system based only on disease severity fails to capture important dimensions of an individual's disease such as the complexity that influences the approach to therapy, level of knowledge and training required for managing the individual case and risk factors that influence likely outcomes. Moreover, the classification should be tailored in a way that can enable the clinician to identify patients with specific forms of periodontitis, and the ability to control and prevent the disease. As a result, during the world workshop, the following classification was established to diagnose patients with PD (see Table 2). From 2018 onwards, clinicians at IUSD have used the classification described in Table 2 [126] to diagnose patients' PD cases.

Table 1: Guidelines for determining the severity of Periodontitis from the American Academy of Periodontology task force report [6]

Periodontal Findings	Slight (Mild)	Moderate	Severe (Advanced)
Probing depths	> 3 & < 5 mm	>= 5 & < 7 mm	>= 7 mm
Bleeding on probing	Yes	Yes	Yes
Radiographic bone loss	Up to 15% of root length or >= 2 mm & <=3 mm	16% to 30% or > 3 mm & <=5 mm	> 30% or > 5 mm
Clinical attachment loss	1 to 2 mm	3 to 4 mm	>= 5 mm

Table 2: Staging intended to classify the severity and extent of a patient's periodontitis status by the American Academy of Periodontology (2018) [124]

Periodontitis stage		Stage I	Stage II	Stage III	Stage IV
Severity	Max. CAL	1 to 2 mm	3 to 4 mm	>=5 mm	>=5 mm
	Radiographic bone loss	Coronal third (<15%)	Coronal third (15% to 33%)	Extending to middle or apical third of the root	Extending to middle or apical third of the root
	Tooth loss	No tooth loss due to periodontitis		Tooth loss due to periodontitis of <=4 teeth	Tooth loss due to periodontitis of >=5 teeth
Complexity	Local	Max PPD <=4 mm Mostly horizontal bone loss	Max PPD <=5 mm Mostly horizontal bone loss	PPD >=6 , Vertical bone loss >-3 mm, furcation involvement Class II & III	Need for complex rehabilitation due to masticatory dysfunction, secondary occlusion trauma (mobility >=2)
Extent and distribution	Localized = <30% teeth involved, and Generalized = < 30%				

In summary, the evolution of AAP classification from 1982 to 2018 to diagnose patients' PD status was described. In this dissertation, a patient dataset from the IUSD clinics was utilized. Dental clinicians at IUSD have used the classification described in Table 1 to diagnose patients' PD diagnoses before the new classification was published in 2018 [124]. From 2018 onwards, clinicians have used the classification demonstrated in Table 2 to diagnose patients' PD status.

2.3 Case-definitions of Periodontal Disease

Case-definitions in epidemiology and population-based studies are a set of criteria used in making decisions about a patient's health status to estimate the prevalence of diseases (for this study, PD) [32, 86]. Case-definitions for PD have been developed by AAP and have been used in the national surveys (NHANES) to determine the prevalence of PD [30, 31, 33, 34, 35]

This case definition of PD determines a patient into severe, moderate or mild/healthy case. The criteria for the different categories are described below.

Severe periodontitis: presence of two or more interproximal sites with ≥ 6 mm clinical attachment loss (CAL) (not on the same tooth) AND one or more interproximal site(s) with ≥ 5 mm periodontal pocket depth (PPD).

Moderate periodontitis: presence of two or more interproximal sites with ≥ 4 mm CAL (not on the same tooth) OR two or more interproximal sites with PPD ≥ 5 mm, also not on the same tooth.

Mild periodontitis: presence of two or more interproximal sites with ≥ 3 mm CAL AND two or more interproximal sites with ≥ 4 mm PPD (not on the same tooth) OR 1 site with ≥ 5 mm PPD [32, 86].

In summary, this chapter highlighted the differences between the PD classifications developed by AAP for diagnosing patients with PD in clinics and case definitions developed to calculate PD prevalence in epidemiological studies. PD classifications described in Chapter 2.2 is mostly used in clinics to diagnose patients' PD status and case definitions described in Chapter 2.3 are used in epidemiological studies and clinical trials to estimate the prevalence of PD. In this dissertation, the case definitions for gingivitis and

periodontitis to automatically determine patients' PD diagnosis from periodontal charting findings were used.

The case-definitions are different from the clinical guidelines used in dental clinics because the clinical diagnosis of periodontitis is based on gingival inflammation, gingival color, measures of pocket depth, radiographic pattern and extent of alveolar bone loss, or a combination of those measures. However, for the case definitions, the diagnosis is solely relying on patients' CAL and PPD values and not intraoral findings, and radiographic bone loss. Therefore, the accuracy and reproducibility of measurements of PPD and CAL are important because case definitions for periodontitis are based largely on either or both measurements and relatively small changes in these values (CAL and PPD values) can result in large changes in disease prevalence [32, 86].

2.4 Calibration of Periodontal Diagnosis and Treatment Planning at Indiana University School of Dentistry

As described earlier, PD is a multifactorial disease caused by the triad of bacterial infection, risk factors (diabetes, cardiovascular diseases, smoking), and host inflammation, it is challenging to formulate a diagnosis based on the classification system [51, 55, 66]. As described in Chapter 2.2, dental clinicians diagnose a patient's PD status by assessing clinical signs of inflammation, location and severity of plaque and calculus, signs and symptoms such as pain or ulceration, periodontal charting findings, and radiographic bone loss information including severity and pattern of bone loss [10]. However, few studies [51, 55, 66, 69] have shown that formulating a diagnosis inherently varies by dentists based on their expertise and experience.

At the IUSD, the department of periodontology and allied dental health is led by Dr. John Vanchit, DDS and Dr. Steven Blanchard, DDS. They have been practicing calibration among dental faculty members and students since 2003 [51]. The IUSD is the only dental school out of 66 dental schools in the US that practices calibration. At IUSD, the effect of consensus training on both faculty members and dental students in the context of periodontal diagnosis and treatment planning was examined [51, 66]. They studied variations in periodontal diagnosis and treatment planning among predoctoral periodontics faculty members after consensus training to compare such variation with those identified in third and fourth-year dental students . Nine cases were given to eighteen faculty members and twenty dental students to diagnose their PD conditions. The authors found that only one patient's treatment plan varied among students and faculty. Moreover, they found that most respondents were able to distinguish clearly among diagnosis of chronic periodontitis,

aggressive periodontitis, and gingivitis. The authors concluded that calibration decreased the variation in periodontal diagnosis and treatment planning among faculty members and dental students.

In summary, in this chapter, the importance of conducting calibration among clinicians due to the multifactorial nature of PD has been highlighted. In this dissertation, clinician-recorded PD diagnoses were utilized from the Indiana University School of Dentistry clinics to determine PD progression over time. Therefore, it is important to highlight that the diagnoses used in this study are calibrated, reliable and accurate. The intensive calibration provided confidence to use the clinician-recorded diagnoses for this research.

2.5 Clinical Course of Periodontal Disease

The clinical course of a disease is defined as the evolution of the disease that has come under medical care and is treated in ways that might affect the subsequent course of events. It is important to know the clinical course of PD to understand how disease initiates, progresses, and factors that contributes towards the progression. So that, dental clinicians can take a step back and assess the risk of the disease initiation or progression for disease prevention. In the section below, findings from the studies [99, 104, 105, 106, 108] that investigated the clinical course of PD is summarized.

Loe et al [64] have studied the clinical course of periodontitis for the first time in 1986 using longitudinal patients' PD data (15 years: from 1970 to 1985) . The authors recruited 480 male tea workers in Sri Lanka and examined their PD status for six times over the course of 15 years. Only 161 participants out of 480 remained by the end of the study. The authors found that 8% of the participants had a rapid progression of PD, 81% had moderate progression, and 11% had no progression of PD.

The second study was performed by Albandar et al [3] in 1991 in which the authors studied the clinical course of calculus formation in men. The authors found that, at age 40 years, all the participants and nearly all of their teeth had developed calculus. They also found that supragingival or subgingival calculus had no impact on loss of attachment. Similarly, they also examined the role of gingivitis in the clinical course of chronic periodontitis using gingival index scores [107, 108]. The authors found that participants who were younger than 40 had a slight increase in attachment loss, and they concluded that gingivitis was one of the risk factors for PD. The authors studied the effect of light smoking on the loss of attachment and teeth in a total of 119 non-smokers and 17 smokers. They

found that current smokers had significantly higher plaque indices, calculus formation, and attachment loss than non-smokers in men aged 35 years or older. The authors confirmed light smoking to be a significant risk factor for PD progression.

Two studies by Schatzle and Wang et al [104, 131] examined the same cohort to find associations between PD and plaque, PD and calculus, PD and gingival indices, and PD and smoking. In the 2009 study by Schatzle et al [104], the authors discovered that increased calculus indices and smoking accelerated the disease progression and that an increased gingival index decreased the rate of regression. A study published in 2017 [99] by Ramseier et al assessed the long-term effect of risk factors such as smoking and calculus on periodontitis and tooth loss in the untreated periodontitis population. The authors found that smoking and calculus were associated with the loss of attachment and progression to advanced PD, and concluded that calculus removal, plaque removal, and control of gingivitis are essential in preventing tooth loss [104, 131]. Other studies also have examined the PD progression and tooth loss by conducting longitudinal studies and they found that smoking, diabetes, and age are causative factors for PD initiation and progression. These studies include [75, 90, 99, 100] [2, 3, 11, 56, 61, 65, 70, 84, 85].

In summary, limited evidence shows temporal associations between the PD risk factors and a change in PD over time. Above described studies discovered that smoking and calculus indices are responsible for disease progression, while increased mean gingival indices and younger age are responsible for the decreased initiation of periodontitis. Researchers were able to find a temporal association, which is difficult to discover through cross-sectional methods.

However, there are several problems with the above described studies. First, a majority of the study cohorts utilized in these studies was from 1969 to 1988, which represents a population from three decades ago. In addition, the subjects were from Sri Lanka, which may not represent the US local population. Second, all of these studies described a major limitation of loss of follow-up visits. For instance, the 2017 study lost 150 patients out of 168 patients (89%) during the follow-up visit. There are more studies needed that can observe patients for a longer period of time to examine the clinical course of PD in the US population. EDR data may be able to provide patients' longer follow-up information on their PD and risk factor status. However, as per the best knowledge, there are only few studies that determined the use of longitudinal EDR data to study PD. In order to examine factors influencing disease progression, treatment outcome or to develop a prediction model, patients' longitudinal data is required. Long-term goal of this dissertation is to determine the factors responsible for PD progression to predict the disease initiation and progression. Therefore, in this study, the quality of the longitudinal EDR data is determined and a cohort of patients' whose disease status changed over time is generated.

2.6 Prediction Models for Periodontal Disease

PD etiology includes the triad of bacterial infection, host inflammation, and risk factors [39]. As described earlier, these risk factors are classified mainly into two categories 1) modifiable, and 2) nonmodifiable risk factors. Modifiable risk factors include systemic diseases, oral hygiene, behavioral factors such as smoking drinking, and diet. Nonmodifiable risk factors are the factors that patients or clinicians have no control for such as age, gender, and genetics. Research has shown that PD could be prevented if the risk factors are controlled [41].

Because PD can be prevented by controlling modifiable risk factors, risk assessment tools and prediction models have been developed for the last three decades to help clinicians with identifying patients who are at high risk of disease. These prediction models include Previser [86, 87, 89], Periodontal Risk Assessment [57, 58], DentoRisk (http://dentosystem.se/en/dentorisk/) [75, 78], and PEMBRA [71, 72, 77]. The periodontal risk assessment tool was created with a simple matrix in Microsoft Excel using IF-ELSE statements and assigned weightage to risk factors based on their scientific evidence [57, 58]. This tool utilized variables, including percent of BOP sites, number of residual periodontal pockets ≥ 5 mm, number of lost teeth, percent alveolar bone loss in relation to the patient's age, interleukin genes, diabetes, cardiovascular disease status, and smoking status. Based on these variables, this tool categorized patients into mild, moderate, or severe risk for PD categories. Similarly, Page et al created a PD risk assessment tool called PreViser [86, 87, 89], using a similar concept of developing a simple matrix; however, PreViser involves more risk factors than periodontal risk assessment. The authors included parameters such as age, diabetes, history of periodontal surgery, BOP, smoking, root calculus, sub-gingival

restorations, pocket depth, furcation involvement, and radiographic bone lesions. Based on these parameters, the tool provides the patient's current PD status (gingivitis or periodontitis), quantifies the disease severity (mild, moderate, or severe) and assesses the risk of future occurrence (1: very low risk, 2: low risk, 3: moderate risk, 4: high risk, and 5: very high risk) [75, 78, 90, 99].

Other PD risk assessment tools have been developed, including UniFe [125], DentoRisk (http://dentosystem.se/en/dentorisk/), and PEMBRA [71, 72, 77], using similar concepts. These models compute risk based on an assessment of current and past findings that have been identified as contributing in some manner to the risk of future disease. Each tool assigns relative risk into one of the arrays of categories that suggest specific treatment approaches to therapy . These tools follow evidence based guidelines in order to deliver treatment recommendations. For instance, if a patient falls in the low-risk category, then the tool recommends treatments such as scaling, root planing, and mouthwash [96].

Studies evaluating effectiveness of these risk assessment tools [71, 72] found that these tools help clinicians rely less on their memory to decide on a treatment plan and help them to improve cognitive function and documentation.

However, despite these advantages, the adoption of these tools is slow because of the following reasons. Studies conducted on validating the performance of these tools in clinical settings [49, 96, 102, 119, 125] have concluded that these tools overestimate or underestimate disease risk for PD. Treatment recommendations generated from these tools are too generic and not patient specific. Therefore, dental clinicians do not trust the performance of these tools, and these tools are not providing new information to clinicians that they do not already know. Some studies have compared performances of different tools

by examining the same patients' PD risk. These studies found that these tools had significantly different risk scores for the same patient. Hence, it was difficult for them to select one tool over the other to assess the patient's risk for PD. They concluded that both tools help predict PD; however, it was necessary for them to use different tools to predict PD because of a lack of integration of all risk factors at a single place and time. They concluded that there is a need for a tool that has all possible risk factors in a single assessment system [73, 102].

In summary, assessing PD risk and approaches to prevent disease initiation and progression has gained significant attention in the last two decades. Multiple risk assessment tools to assess PD diagnosis, risk, and treatment recommendations have been developed. These tools help clinicians with their thought processes, increase their self-efficiency, and improve psychological parameters, positive cognitive behavior, and documentation. Despite this effort, the use of these risk assessment tools remains low and PD is still prevalent. While these tools help clinicians identify risk factors, they fail to identify which factor is driving the risk and how to control it, which would enable a shared decision. They either overestimate or underestimate disease risk, and treatment recommendations provided by these tools are similar for any patients that fall under the same category despite providing patient specific treatment recommendations. As a result, while these tools assist clinicians in determining their patients' risk factors, they fail to assist clinicians in stopping the initiation and progression of the disease in their patients. Recent studies have demonstrated that, while these tools were developed based on established scientific evidence, scant knowledge exists as to how these tools can assist clinicians in changing patient behavior to enhance preventive care. Moreover, there has

been a change in patients' disease profiles and lifestyle behaviors over the last two decades, which are not considered by these tools. EDR has the potential to provide patient data that can represent the patient's population and help with assessing patient's disease risk accurately. However, if the quality of the EDR data is poor then the output generated by these models will be poor as well. Therefore, in this dissertation, the quality of the EDR data was evaluated before its use for research and to develop prediction models.

2.7 Electronic Health Record: Secondary Use of Electronic Health Records

The increased availability of patient care data through EHRs offers an opportunity to conduct research studies, such as observational studies, clinical research, and quality improvement [25]. Since EHR data is not stored for research purposes, but rather for clinical care purposes, it can provide more generalizable results and may be able to provide a diverse population that is representative of actual patients [23, 25]. Similarly, EDR data could be utilized to characterize present patient population risk factors and disease profiles with up-to-date information, compared to prospective studies that were performed two decades ago [120]. Additionally, this data could be utilized to study PD treatment outcomes and build a PD prediction model that represents the present population [25].

Evaluating the quality of the EHR data is critical before its use for research:

Due to the increased availability of various statistical methods and tools, many researchers are working on developing prediction models utilizing EHR data. However, the use of these models is not well understood. Despite the immense advantage of EHRs, which are capable of providing up-to-date data representing the current patient population, if the quality of the data fed into these models is poor, then the validity and reproducibility of results will be affected and provide spurious outcomes. Therefore, assessing EHR data quality before its use for any research studies is critical. In the section below, detailed description about the data quality measures used in this dissertation is described.

Concordance:

According to Weiskopf et al [133] data are considered concordant when there is agreement or compatibility between data elements. This may mean that two elements recording the same information for a single patient have the same value, or that elements

recording different information have values that make sense when considered together. It helps in answering the question: “Is there agreement between elements in the EHR, or between the EHR and another data source?” Measurement of concordance is generally based on different fields contained within the EHR, and/or it also includes information from other data sources.

Completeness

Completeness is the most commonly assessed dimension of data quality [133]. Completeness refers to whether or not a true fact about a patient was present in the EHR. Most studies [36, 60, 67, 133, 134] use the term completeness to describe this dimension, in addition to data availability or missing data. They have compared their EHR data with the gold standard. The gold standard includes paper records [16, 67] information supplied by patients [134], a review of data by patients clinical encounters with patients [45, 46], information presented by the trained physician, and alternative data sources. [101]. Some studies [101, 109] measured their data completeness by comparing distributions of occurrences of certain elements between practices or with nationally recorded rates.

Plausibility, currency, and granularity:

In the existing studies [36, 67, 80, 101, 133, 134], plausibility was calculated by comparing the data study results with general medical knowledge or information. The data quality measure “Currency” was often referred to as timeliness or recency. Data is considered to be current if they are recorded in the EDR within a reasonable period of time following a measurement or, alternatively, if they were representative of the patient state at a desired time of interest. This is usually checked via data entry logs as shown in studies [80, 101, 133].

Granularity measurements help define EDR data as too general or too specific as shown in studies [80, 101, 133].

Among the studies [36, 67, 80, 101, 133, 134] evaluating the quality of EHR data, most have raised concerns about poor data quality. For instance, Skyttberg et al [114] studied the effects on data quality of vital signs on three different types of documentation practices. The authors found that the vital signs documented in Swedish emergency care EHRs cannot generally be considered as high enough quality, and they discovered low completeness and currency of this information. Martins et al [67] measured variation among four different EHR systems documentation locations versus the "gold standard" manual chart review for risk stratification in patients with multiple chronic illnesses. The authors found that patient information was recorded in inappropriate EHR locations 30% of the time. Similarly, Singer et al [113] studied problem list completeness related to chronic diseases in EHRs. The authors found high variability and generally low quality of problem lists related to seven common chronic diseases.

In dentistry, limited studies have evaluated the quality of the EDR data and developed process measures for the quality improvement program [15, 54, 60]. These studies estimated the percentage of patients who received fluoride varnish, dental sealants, and treatment provided to periodontal disease patients with diabetes. Although, the goal of these studies was not to evaluate the quality of data reported in the EDR, one of the studies mentioned that narrative notes provided more information regarding treatments compared to the procedure codes [54].

Three recently published studies evaluated the completeness and correctness of patient demographics, dental treatments [120] clinical documentation of patient's diagnosis

of alveolar osteitis [60], and medication documentation in the EDR in student clinics [15]. Thyvalikakath et al [120] examined the completeness and correctness of patient demographics, insurance status, and variables required to examine the survival analysis of root canal treatment, and posterior composite restorations from 99 private dental practice EDR data. They found excellent data quality of patient demographic, and variables needed to perform root canal treatment and posterior composite restorations. Levitin et al [60] examined the completeness of alveolar osteitis information by comparing population-based prevalence and frequency of corresponding items in the student documentation. They found a wide discrepancy and concluded that more attention to clinical documentation skills is warranted in dental student training. Burcham et al [15] examined the completeness of medication documentation in the EDR and found that only 1.1% of the records as "were completely documented". Although this study found very fewer less complete records, the criteria used that defines "complete record" were stringent. These criteria include 1) proper medication name, 2) medication dose/frequency, 3) oral side effects, and 4) correct spellings. It is humanly impossible to correctly document each medication name, and dosage without any grammatical or spelling error. This study also reported that patient's' medication and its associated medical conditions were reported fragmentary in the EDR. Both of these studies showed that dental clinicians report patient's' disease signs, symptoms, and medication information up to a certain extent but not necessarily in the same data field with spelling and grammatical errors. Moreover, they tend to document signs and symptoms which are essential to diagnose a condition but may not report the diagnosis; or they report diagnosis in the clinical notes but do not use diagnostic codes to document diagnosis [15, 60] . Acharya et al [1] performed a study assessing regional PD

prevalence in Wisconsin. The authors found that 45% of the Wisconsin population had any kind of periodontitis, which is consistent with the national periodontitis prevalence estimation [1]. Despite this effort, as per the best knowledge, no studies have examined the quality of PD data stored in the EDR.

2.8 Common Data Elements' Between Electronic Health Records and Prospective Studies

In addition to assessing the secondary use of the EHR data described in the previous chapter, another way is to examine the quality of the EHR data is through identifying common data elements. Common data elements are defined as metadata information that is of interest or relevance in a specific research domain [14]. Common data elements are the variables which are collected in prospective studies for research purposes and also reported in EHR while providing regular patient care. While designing prospective studies, researchers decide the types and depth of variables to be collected during the study. Since in EHRs patient information is collected for clinical care purposes and not for research purposes, clinicians may or may not collect patients' all disease related variables in a granular fashion [23, 25]. Therefore, examining common data elements between EHRs and prospective studies help in determining up to what extent information is available in the EHR regarding patients' PD status and risk factors.

Some studies [14, 26, 27, 62, 103, 118] have examined the common data elements between EHR data and information collected in prospective studies. Liu et al [62] conducted a data mapping study in which authors examined the availability of variables regarding comprehensive taxonomy for general dentistry and variables used in dental practice-based research network studies. Authors found that, 33% of the variables used in practice based research network studied matched and were available in the EDR which include data about dental anatomy, medications, and items such as oral biopsy and caries.

Similarly, Bruland et al [14] have investigated the common data elements for secondary use of EHR data for clinical trial execution and adverse drug reporting. They

generated case report forms for total 23 clinical trials in different disease areas and sorted variables generated in these trials. They examined if these variables were available in the EHRs. They found that the variables related to reimbursement are frequently available, while, more specified variables were not frequently available . Doods et al [26, 27, 28] have examined the presence of clinical trial data elements within existing EHR systems and found a broad range of coverage between 13 to 70%.

A study by Violan et al [129] compared information provided by EHR data with a population health survey to estimate the prevalence of several health conditions (cardiovascular diseases, diabetes, asthma, anemia, etc.). The authors found that the prevalence of self-reported multimorbidity was significantly higher in health survey data among younger patients while prevalence was similar in both data sources for elderly patients. Self-report provided more sensitive data to identify symptoms-based conditions . From the existing studies, it's conclusive that there is wide variation among the findings from the results and further research is warranted. While there is some approach taken in medicine, very limited studies in dentistry have examined the common PD data elements between EDR and clinical trials.

2.9 Phenotype Patients Disease Status Utilizing Electronic Health Record Data for Research

As described in Chapter 2.6, EHR data offer the opportunity to conduct a large-scale population-based research studies quickly while minimizing cost, and time [25]. However, since this data is collected for patient care and not research, they represent clinicians' observations and actions on the patient rather than the patient himself. Also, often times the variables of interest/disease concepts are placed in different sections of the EHR. For instance, patient self-reported diabetes status is in multiple narrative notes, structured format, or in scanned questionnaire. Moreover, patient may not self-report of his diabetes status, however, may report taking medications for it. Therefore, utilizing all possible sections of EHR is important when using this data for research. Also, data recorded in narrative notes such as clinical notes, progress notes are difficult to extract in a reliable format to be used for research purpose. Therefore, before utilizing this data, it is important to transform the raw EHR data to a form that is useful for clinical research.

The method of phenotyping maps the master data to intermediate states like inferred clinical conditions that are then used in research [94]. In medicine, many studies have developed phenotype methods to map the raw EHR data to patients' disease state by utilizing multiple fields [37, 94, 111]. These studies have used multiple EHR data fields such as medical history, structured diagnostic codes, laboratory reports, and medication histories to phenotype diabetes information [7, 117], cardiovascular diseases [76], and mental diseases such as depression [48]. Various approaches have also been utilized to phenotype patients' disease status. For example, Pathak et al used a set of rules or queries that assert disease

status using the raw EHR data [94], or using machine learning methods to phenotype diseases as described in a review study [112].

There are only few studies [1, 20] conducted in dentistry in which researchers have utilized EDR data to phenotype patients' PD status . Acharya et al [1] examined regional epidemiology of periodontitis using EDR in Marshfield clinical research foundation. Authors phenotype patients' periodontitis status using the master data from the EDR . Similarly, Chatzopoulos et al [20] utilized EDRs to assess periodontitis prevalence and its association with systemic diseases. Authors manually reviewed 5000 patients radiographs and dental record to determine periodontitis prevalence.

These studies provided new knowledge on potential of utilizing EDR data to assess periodontitis status. However, authors they did not assess the quality of the EDR data before using the data for research. Moreover, they manually reviewed records which may not be feasible on a bigger sample of population. And most importantly, these studies only have examined the prevalence of periodontitis and did not examine the prevalence of gingivitis which is a precursor of periodontitis.

In summary, there has been an extensive effort on using EHR data for research purposes and advance statistical methods to develop prediction models for risk assessment. Moreover, many studies also discovered various advance computational methods to extract and convert EHR information in an analyzable format for research and patient care purposes. Moreover, in medicine some studies have assessed the quality of the EHR data and developed methods and framework to assess EHR data quality (see Chapter 2.7.1).

Despite this extensive efforts, limited research has been conducted in order to assess the quality of the EDR data. There is no study exist that can help in determining the quality

of the PD and its risk factor information stored in the EDR and answer the question “Can EDR provide patients’ accurate PD variables and its risk factor information?”.

Leveraging Electronic Dental Record Data for Research

Due to the high adoption of EDR since the last decade, many studies [20, 21, 1, 20, 22, 44] have utilized EDR data to conduct research studies. For example, Chatzopoulos et al [21] studied implant and root canal survival rates and factors associated with treatment outcomes by utilizing electronic dental records . They utilized 13,434 patient records and found that the overall survival rate was significantly higher for implant therapy compared to root canal treatment. They also found that age and anxiety worsen the treatment outcome of these patients. The authors also published a paper on systemic medical conditions and periodontal status in older individuals . In this study the authors used 2,163 patients’ records to assess the prevalence of PD and its association with systemic diseases. They found that self-reported tobacco use and diabetes were significantly associated with moderate and severe bone loss. Surprisingly, they also that patients who had joint replacement, past use of steroids and acid reflux/GERD had less severe PD compared to others. Recently, Laske and Opdam et al [22] published a study on risk factors for dental restoration survival . Authors utilized practiced based research network dental data to identify risk factors that affect dental restoration survival. Acharya et al [1] have utilized integrated electronic dental and medical record data to estimate the prevalence of periodontitis in the Wisconsin population. They found that around 45% of the Wisconsin population is suffering from periodontitis which is equivalent to the national prevalence reported in the NHANES study [31].

At the IUSD, the dental informatics program led by Dr. Thankam Paul Thyvalikakath have published several studies in which the team has leveraged EDR data for dental research. Thyvalikakth et al recently published a study [120] on "Leveraging EDR data for clinical research in the National Dental Practiced Based Research Network Practices". Authors recruited 99 dental practices to evaluate the data completeness and correctness for the variables needed to assess survival analysis of two dental procedures 1) root canal treatment, 2) posterior composite restoration. They found nearly 100% completeness of data, and 80% of correctness of data. This study demonstrated the feasibility of using EDR data for dental research. The authors established the groundwork for a learning health system that will enable practitioners to learn about their patients' outcome by using their own practice data. The dental informatics group members at IUSD have developed text-mining pipelines to extract patients' smoking [92, 93], and cardiovascular diseases [91] from the free-text of EDR data . In the smoking project, the dental informatics team leveraged EDR to extract patients' detailed smoking information based on smoking intensity which may not be available in the EHR [93]. The authors trained three machine learning algorithms using the training set of 2,176 patient records and tested performance on 1,120 records. They achieved excellent performance for automatically classifying patients based on their smoking intensities and concluded that 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 (see Table 3) [92, 93]. The dental informatics team also compared what information is reported in the EDR regarding cardiovascular diseases with patients linked dental and medical records [91]. They achieved excellent agreement in extracting patients' CVD information from the EDR (see Table 4).

Low agreement between self-reported EDR data and physician-diagnosed EMR data [93] was observed. The team members at IUSD have also leveraged EDR data to determine the differences in medication usage by dental patients' age, gender, race, and insurance status. It was observed that 12 to 14% of patients were between 18 to 54 years old reported taking opioid agonists and Selective serotonin reuptake inhibitors medications. This study highlighted the importance of taking medical and medication history regardless of their age to avoid adverse events during dental care [132], (unpublished data).

In summary, it is now evident that EDR and EHR data provides promising potential to conduct clinical research. Many research questions can be answered by using the EDR and EHR data. However, in order to utilize this data to its optimum potential, it is important to utilize advance informatics methods. For example, as described in Chapter 2.7.2, 2.7.3, phenotype algorithms need to be developed to determine patients' accurate disease status because information can be present in multiple sections of the EHR. This chapter is important for this dissertation, because, first, it shows the important of evaluating the quality of the EHR and EDR data before its intended use. As flawed data results into flawed outcome. Second, this chapter shows the importance of informatics methods to optimally use EHR data for clinical research. Without informatics methods, manual review would be required to use this data that is time consuming, expensive, and error some. Last, this chapter shows the overall work dental informatics work done in the dental community, and at the dental informatics led by Dr. Thankam Paul Thyvalikakath that demonstrates the confidence and feasibility of successfully achieving the proposed aims.

Table 3: Performance of Support Vector Machine to classify patients based on their smoking intensity

Smoking status	Precision (%)	Recall (%)	F-measure (%)
IntS	95	98	96
IS	90	88	89
PS	89	89	89
LS	87	87	87
SUI	76	86	81
HS	90	44	60
Average	89	82	84

Abbreviations: IntS: Intermittent Smoker, IS: Intermediate Smoker, PS: Past Smoker, LS: Light Smoker, SUI: Smoker with Unknown Intensity, HS: Heavy Smoker

Table 4: Performance of the CVD Extractor for encoding CVD condition concepts, CVD procedural concepts, CVD negation attributes, CVD experiencer attributes, CVD temporality attributes, and CVD severity attributes

CVD concepts and attributes	Precision (%)	Recall (%)	F-score (%)
Overall	98	76	85
CVD condition concept	100	64	78
CVD procedural concept	98	62	76
CVD negation attribute	94	88	91
CVD experiencer attribute	100	100	100
CVD temporality attribute	98	62	76
CVD severity attribute	100	83	90

3: Aim 1 Methods

Aim 1

Determine the completeness and concordance of patients' periodontal disease diagnosis recorded during comprehensive oral evaluation in their electronic dental record data.

Hypothesis

We hypothesize that the completeness of PD diagnosis information recorded in the EDR data will be at least 80% complete. The value of completeness was set at 80% because, at the Indiana University School of Dentistry (IUSD), dental students are mandated to record periodontal findings such as clinical attachment loss, periodontal pocket depth, and bleeding on probing that derived from periodontal charting. It is hypothesized that the concordance between PD diagnosis determined from periodontal charting and clinician-documented diagnosis will be at least 90%. The value of 90% was set for concordance because all patients are required to have complete PD diagnosis and findings documentation while receiving a comprehensive oral evaluation in the EDR. Moreover, periodontal findings utilized to determine PD diagnosis from charting data were the same as the findings used by the clinicians to determine PD diagnosis.

To achieve this aim, the following steps were performed. First, a patient cohort was generated from the Indiana University School of Dentistry (IUSD) clinics. Second, patients' PD diagnosis from the periodontal charting findings were generated (bleeding on probing, clinical attachment loss, periodontal pocket depth). Next, clinician-documented PD diagnoses were retrieved from the "diagnosis" section of the periodontal evaluation form. Next, the data quality measure "completeness" was measured to assess the percentage of completeness of clinical variables such as patient demographics, periodontal

charting, and PD diagnosis in the periodontal evaluation form. Last, concordance between PD diagnosis derived from periodontal charting findings and clinician-documented diagnosis recorded was determined. The results of this study would contribute to determining the completeness of PD diagnosis for patients who underwent COE. This study result would also contribute to determining the appropriateness of utilizing EDR data for research, to assess treatment outcomes, and to predict PD.

3.1: Data Source & Study Design

Data Source

EDR (axiUm®-EXAN) data from the IUSD pre-doctoral clinics to conduct this study. The EDR system was implemented at IUSD in the year 2005 was utilized. Patients who visited IUSD and IUSD satellite clinics and received at least one comprehensive oral evaluation (COE) between January 1, 2009, to December 31, 2014 were retrieved. The patients' visit information that may fall outside this time period were also included in this study. For example, if a patient received COE in 2010 and received treatments in 2007 and 2015 then, information from 2007 and 2015 would also be included in this study. This time period was selected because it was aimed to compare prevalence generated in a recently conducted epidemiological study [33]. This study has estimated the prevalence of periodontitis in the United States between 2009 and 2014. Selecting this similar timeline would allow an effective comparison between this study results and with the results of epidemiological study [33].

Study Design

The design of this study is a retrospective cohort study.

3.2: Inclusion & Exclusion Criteria

Inclusion criteria

This study dataset included patients who underwent COE between January 1, 2009, to December 31, 2014, and who were 18 years or older during the time of their first completed COE procedure. Four main data tables were utilized in axiUm®: 1) patient demographics (date of birth, gender, race, insurance), 2) procedure codes (treatments provided), 3) periodontal charting (CAL, PPD, BOP), and 4) periodontal evaluation form (clinician-recorded PD).

Exclusion criteria

Patients who were less than 18 years old and who did not receive COE during the study time period (January 1, 2009, to December 31, 2014) were excluded.

3.3: Study Variables

Demographics

Patients' demographic information such as date of birth, gender, and race/ethnicity, was retrieved. Patients' age was calculated from their date of birth information recorded during their first COE examination.

Insurance status

Patients' insurance information was also retrieved. Patients' insurance information was recorded in three structured categories: 1) government-insured, 2) privately insured or 3) self-paid.

Periodontal Examination Findings

Periodontal examination findings are documented within two sections in the EDR: 1) periodontal findings in the periodontal charting section, and 2) clinician-recorded diagnosis in the periodontal evaluation form. From the periodontal charting section, the variables used to diagnose PD are collected. These variables are described in the following sections.

Bleeding on probing (BOP):

The BOP score is used to diagnose the presence of gingival inflammation and to diagnosis gingivitis. The BOP score is also a predictor for the progression of PD. More information on BOP is present in the background section (see Chapter 2.2). When BOP is present, to document a patient's BOP site, dental clinicians usually click on the tooth surface where BOP was present. Each tooth has a total of six surfaces which include mesiobuccal, midbuccal, distobuccal, mesiolingual, lingual, and distolingual (see Chapter 2.1). After

clicking, the affected site will have an automatic red dot or “B” in the charting interface. A screenshot of the BOP charting in axiUm® is present in Figures 2, 3.

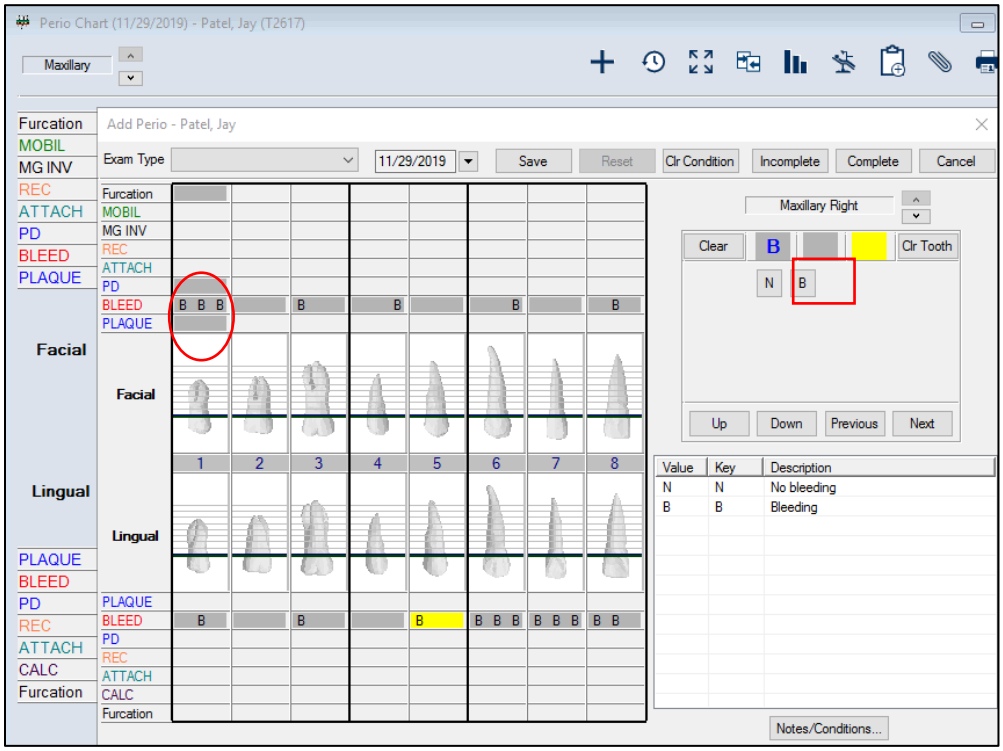


Figure 2: Documenting bleeding on probing findings in the axiUm® electronic dental record’s periodontal charting module

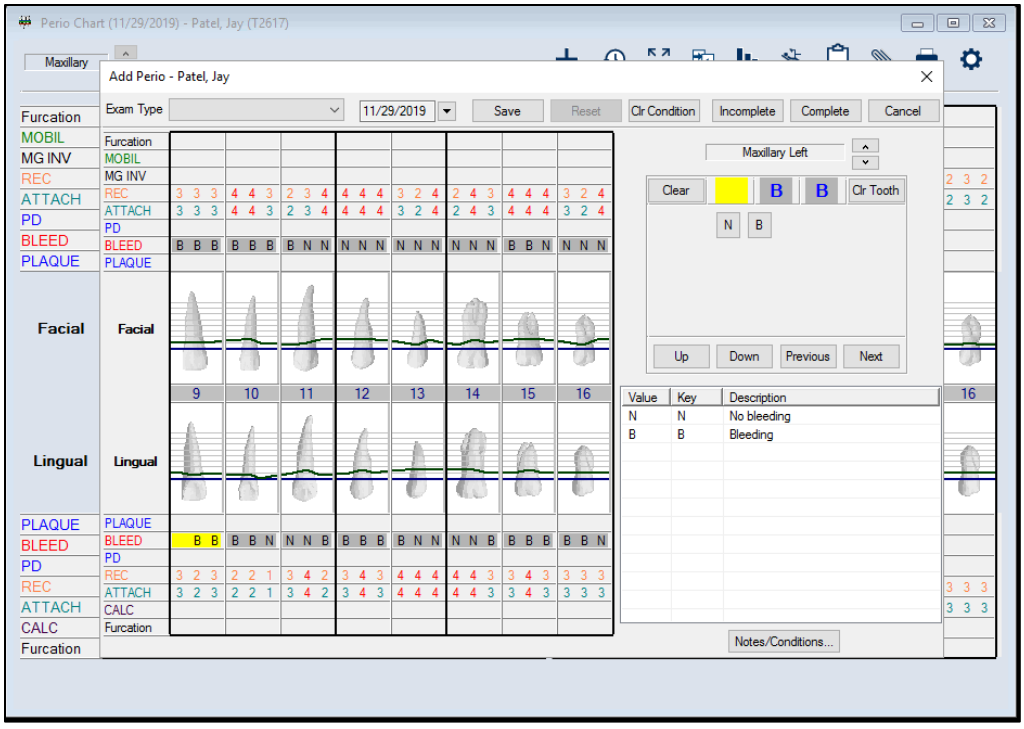


Figure 3: Documented bleeding on probing information in the axiUm electronic dental record's periodontal charting module

Periodontal pocket depth (PPD)

PPD is the distance between the free gingival margin and the bottom of the gingival sulcus. Dental clinicians measure PPD at each of the six sites of the tooth which include mesio-buccal, mid-buccal, disto-buccal, mesio-lingual, lingual, disto-lingual. As shown in Figures 2 & 3, clinicians enter PPD information in millimeters per tooth-site in axiUm®.

Clinical attachment loss (CAL)

CAL is the distance between the gingival recession margin and pocket depth. In axiUm, based on the entered PPD and gingival recession, CAL is automatically calculated. CAL value is calculated by adding the value of PPD and gingival recession (see Figure 4). For example, if a patient has 4 mm of PPD and 3 mm of the gingival recession on the mesiobuccal site of tooth 4 then, the CAL value is 7 mm (4 mm + 3 mm).

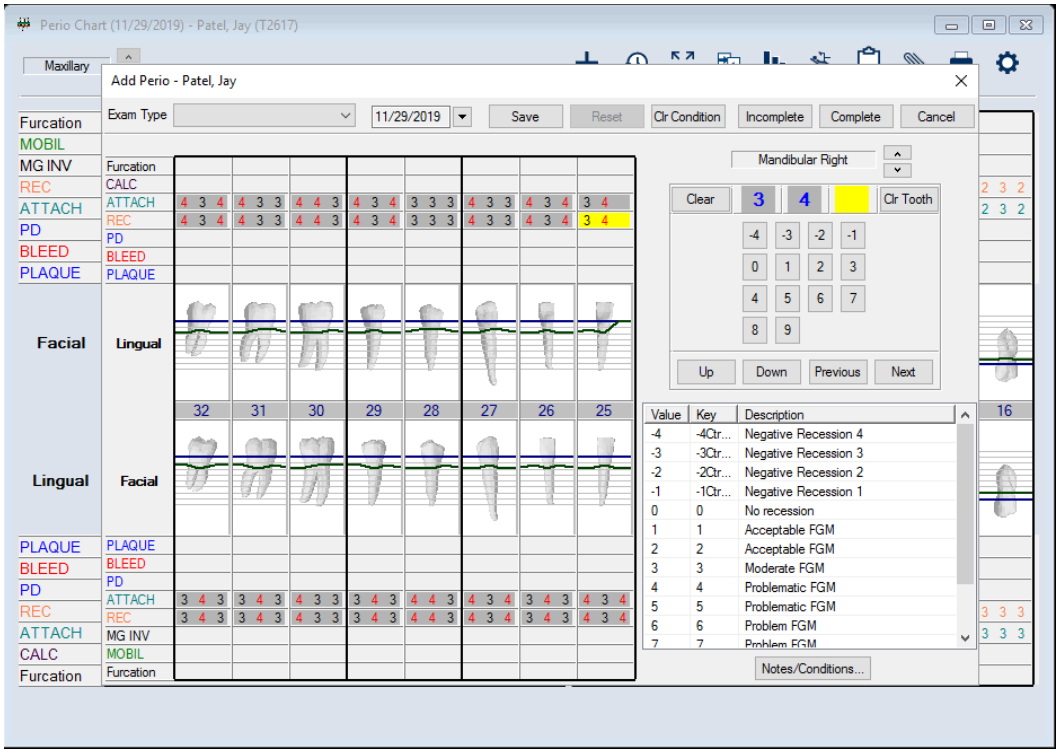


Figure 4: An example screenshot of documenting gingival recession and attachment information in the axiUm® electronic dental record's periodontal charting module

An example of a complete documented periodontal chart that includes complete documentation of CAL, PPD, and BOP is shown in Figure 5.

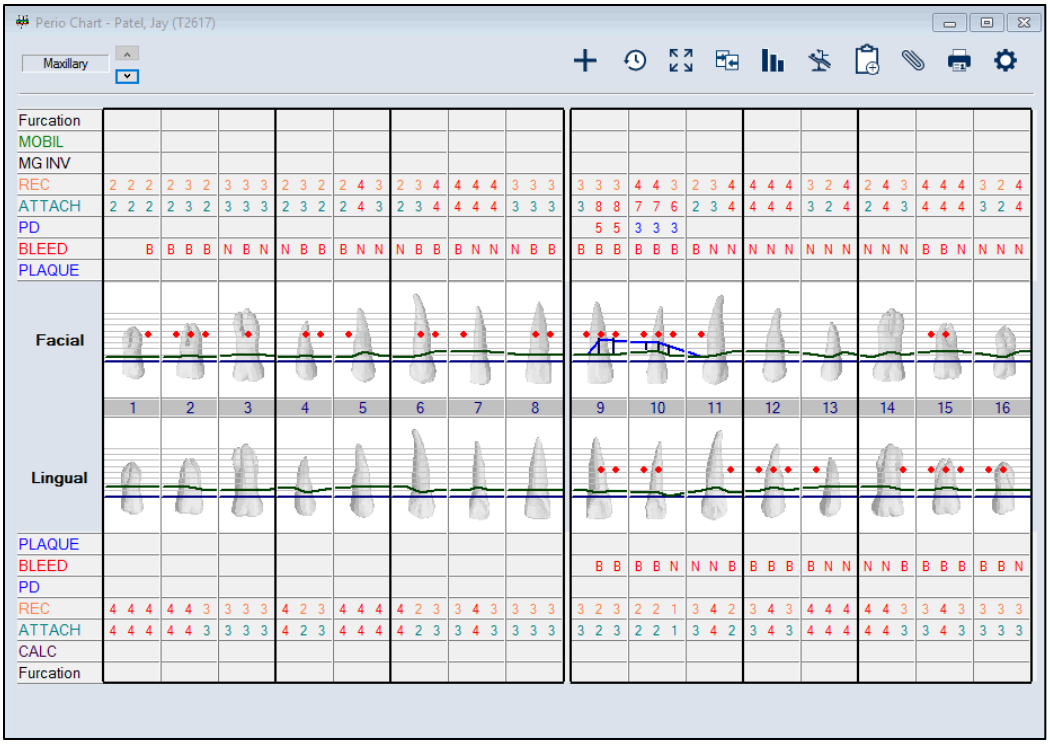


Figure 5: An example of a patient's complete periodontal charting documentation in axiUm® electronic dental record's periodontal charting module

Clinician-recorded PD diagnosis

Based on CAL, PPD, BOP, intraoral examination, and radiographic findings, clinicians record their patients' PD diagnoses in periodontal evaluation forms in a separate section "diagnosis" within axiUm.

Treatment histories

We included patients' COE (procedure code: D0150), periodic oral evaluation (POE) (procedure code: D0120), periodontal maintenance (PM) (procedure code: D0140), and periodontal re-evaluation (PRE) (procedure code: D0450). During thee COE, dental clinicians document patients' complete medical history, dental history, intra oral findings, extra oral findings, diagnosis, odontogram findings (the chart that depicts hard and soft tissue findings), periodontal charting, oral cancer evaluation, and treatment

planning. During the POE, dental clinicians assess any changes in the patient's dental and medical health status following a previous periodontal or restorative treatment based on COE. This includes oral cancer evaluation and periodontal screening.

3.5: Clinical Workflow and Documentation of Patient Care

During COE and routine dental care visits, dental students perform the intraoral examination and record their patients' medical history, dental history, dental diagnosis, treatment plan, hard and soft tissue findings, radiographic findings and prognosis. Regarding periodontal findings, they document patients' BOP, PPD, CAL, mobility, and furcation involvement in the periodontal charting section in the EDR (see Figures 2, 3, 4). These findings are then checked by periodontal clinical faculty. Upon review, students use these periodontal charting findings, intra oral findings such as the extent and pattern of gingival and periodontium, and radiographic bone loss to diagnose PD. PD diagnosis is typically recorded in the "diagnosis" section in the periodontal evaluation form.

Therefore, in the EDRs, PD diagnosis could be obtained from three different fields in the EDR: 1) in the "periodontal charting (findings)", 2) in the "diagnosis" section of the periodontal evaluation form, and 3) structured diagnostic codes from the treatment planning module (only in use since 2014). In this study, two fields (periodontal charting findings, and diagnosis section of the periodontal evaluation form) are used to determine patients' PD diagnosis. The third field (structured diagnostic codes) was not used because this module was only used from 2014 onwards, which falls outside of the study period (January 1, 2009, to December 31, 2014).

3.6: Data Pre-processing

3.6.1: Creating Study Subsets

First, patient demographics information, periodontal evaluation forms, treatment histories, and periodontal charting data were received in tab-delimited text files (source dataset/master dataset) from the axiUm® Oracle database. Information present in the source dataset is described in Table 5. From this master dataset, four subset text files were created to conduct this study (see Figure 6). These variables include demographics (date of birth, gender, race), insurance status, periodontal charting findings (CAL, PPD, BOP), clinician-documented PD diagnosis, and treatment history. To create subsets, a computer algorithm, "*Subset Extractor.py"* using inbuilt Python (a computer programming language) functions, was developed. Table 6 shows detailed information about each of these functions used to develop *Subset Extractor.py*. More information on the development of the program is present in Appendix (see page 189). The content among these four subsets of text files is described as follows.

1) Patient Demographics: Retrieved patient demographics information such as date of birth, gender, race, and insurance status recorded during their first COE.
2) Periodontal charting findings: Retrieved patients' BOP, PPD, CAL which were recorded during their first completed COE.
3) Periodontal evaluation forms: Retrieved patients' periodontal evaluation forms information from their first and consequent visits. For this specific aim, patients' PD information recorded during their first completed COE was included.
4) Treatment history: Retrieved patients' completed COE, POE, PM, and PRE information.

Table 5: Final study dataset received after querying the database and variable of interest extracted from the study data using subset extractor computer algorithms

Information type in the EDR	EDR data	Variable of interest
Demographics	date of birth, race, gender, insurance status, chart IDs	date of birth, race, gender, insurance status
Periodontal charting	clinical attachment loss, periodontal pocket depth, bleeding on probing, furcation, mobility, plaque score, calculus score	clinical attachment loss, periodontal pocket depth, bleeding on probing
Periodontal evaluation form	periodontal condition, extent and pattern, periodontal diagnosis, etiology, prognosis, number of teeth in each quadrant, percentage of number of clean teeth in each quadrant.	Periodontal condition, extent and pattern, periodontal diagnosis, etiology
Treatment history	procedure codes	Comprehensive oral evaluation, periodic oral evaluation, periodontal maintenance, periodontal re-oral evaluation

Table 6: Description of python functions and library used in developing computer algorithms to create study subsets, and automatically calculate the bleeding on probing score to diagnose gingivitis

Python functions and libraries	Description
Readlines ()	Returns list containing the lines.
List ()	Returns a mutable sequence list of elements. If no paraments are given (e.g. list ()), then it creates an empty list.
Map ()	Returns a list of the results after applying the given function to each item.
Replace ()	Returns a copy of the string where all occurrences of a substring is replaced with another substring (depending on the given input).
Len ()	Returns the length of the string.
Search ()	This is a regular expression for describing a search pattern.
Re()	Regular expression function to search for keywords and the pattern of writing.

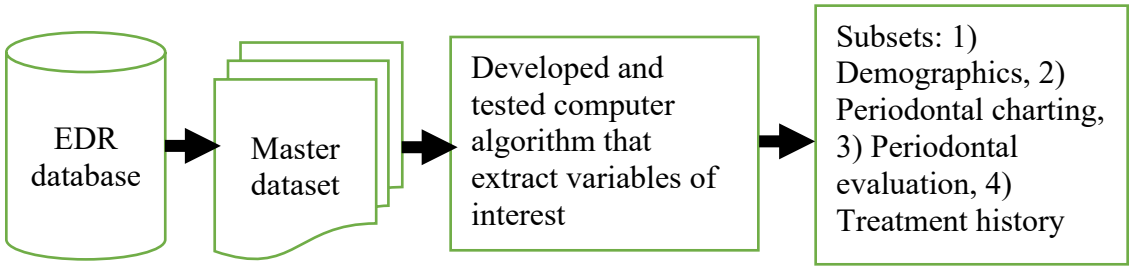


Figure 6: Creating study subsets (demographics, periodontal charting, periodontal evaluation form, and treatment history) from master dataset

3.6.2: Converting Individual Patient Record into Text Files

A computer algorithm (*TSV to Individual Text File.py*) that converts and stores each patient's periodontal charting information in individual text files was developed. Each patient's charting information was stored in a separate text file because of two reasons: 1) to determine PD diagnosis based on charting findings, and 2) to track disease progression over time from recorded dates. Each converted text file contained patients' ID, charting date, periodontal findings (CAL, PPD, or BOP), tooth number, and tooth sites (mesiolingual, mesial, mesiofacial, distal, distolingual, distofacial). The text file format was "Patient ID_visit_Date_of_Visit.txt", which helped in tracking patients' PD diagnosis during each visit. By the end of this step, individual text files were converted by unique patient ID & visit dates for further processing and analysis. The source code of this computer algorithm is described in Table 45 (See Appendix section on page 193).

3.7: Develop and Test an Algorithm to Diagnose Patients' Gingivitis Status from Periodontal Charting Findings.

3.7.1: Gingivitis Diagnosis Criteria

As described in Chapter 2.2-periodontal disease diagnosis, there are no universally accepted criteria to diagnose a patients' gingivitis status. A recently published study on the diagnostic consideration and case definition of plaque induced gingivitis claimed that the BOP score should be the standard criteria for gingivitis diagnosis because it is an objective, reliable, and accurate measure [126]. As a result, the BOP score was used to diagnose patients' gingivitis status. A BOP score is assessed as the proportion of bleeding sites (dichotomous yes/no evaluation) when stimulated by a standardized (dimensions and shape) manual probe with a controlled (~25 g) force to the bottom of the sulcus/pocket at six sites (mesiobuccal, buccal, distobuccal, mesiolingual, lingual, distolingual) on all present teeth. The BOP score is used for classifying patients into healthy, localized or generalized gingivitis case. According to this classification criterion [126], when the BOP score is 1) less than 10%, it is diagnosed as a healthy case, 2) from 10 to 30%, it is diagnosed as a localized gingivitis case, and 3) more than 30%, it is diagnosed as a generalized gingivitis case.

3.7.2: Developing the Algorithm Using Python

A computer algorithm was developed (*Gingivitis_Diagnoser.py*) that classified patients' gingivitis status into healthy, localized, or generalized gingivitis cases based on the BOP score. *Gingivitis_Diagnoser.py* consisted of various Python inbuilt functions, variables, and rules (see Table 6). First, this algorithm created a variable, *"Total_Sites",* and determines the total number of sites present in a patient's dentition. To calculate the BOP score*,* the value for the number of total sites is mandatory to use it as a denominator. To calculate the total sites, the algorithm evaluated the content in each text file using the indexing approach. It examined each tooth number in the file and saves it in a temporary variable *"Number of Teeth"*. To calculate the BOP score, the total number of sites based on the total number of teeth is required (BOP score = total number of BOP sites / total number of sites). Next, the *Gingivitis_Diagnoser.py* counts one tooth number only once to prevent counting the tooth number twice. For example, in the text file, typically, the tooth numbers are recorded twice, one recording for the buccal side of the tooth, and one for the lingual side of the tooth. Therefore, both buccal and lingual findings belong to the same tooth number. After counting the total number of teeth, the algorithm saves this information in the *"Number of Teeth"* variable. Then, the total number of teeth is multiplied by six using the multiplication function to obtain the total number of sites. The total number of sites in the patient is stored in a variable *"Total_Sites"*. Similarly, the number of sites with positive BOP is stored in the *"Number of Bleeding on Probing Sites"* variable. Next, the BOP score is calculated by dividing the number of BOP sites by the total number of sites. The value of the BOP score is stored in a variable *"BOP_Score"*. Lastly, the rules (see Table 7) to determine the patient's gingivitis diagnosis were developed.

Table 7: Rules to determine patients' gingivitis diagnosis based on the bleeding on probing scores

IF (number of bleeding on probing sites / total_sites) * 100) < 10: THEN 'No_Gingivitis'
IF (number of bleeding on probing sites / total_sites) * 100) >= 10 AND <= 30: THEN 'Localized_Gingivitis'
IF (number of bleeding on probing sites / total_sites) * 100) > 30: THEN 'Generalized_Gingivitis'

This algorithm placed each text file into one of the three folders 1) No gingivitis, 2) localized gingivitis, or 3) generalized gingivitis. For example, if a patient had 48 BOP sites present in his record and he had 28 teeth present (determined from the value of CAL), then this patient's BOP score would be 29% [48 (total number of BOP sites) /28*6 (total number of sites) = 29% (the BOP score)]. Therefore, according to the diagnostic criteria, this patient was diagnosed as localized gingivitis case. The illustration of encoding each patient in healthy, localized or generalized gingivitis cases is described in Figure 7. The source code of this computer algorithm is described in Table 46. (See Appendix section on pages 194-197).

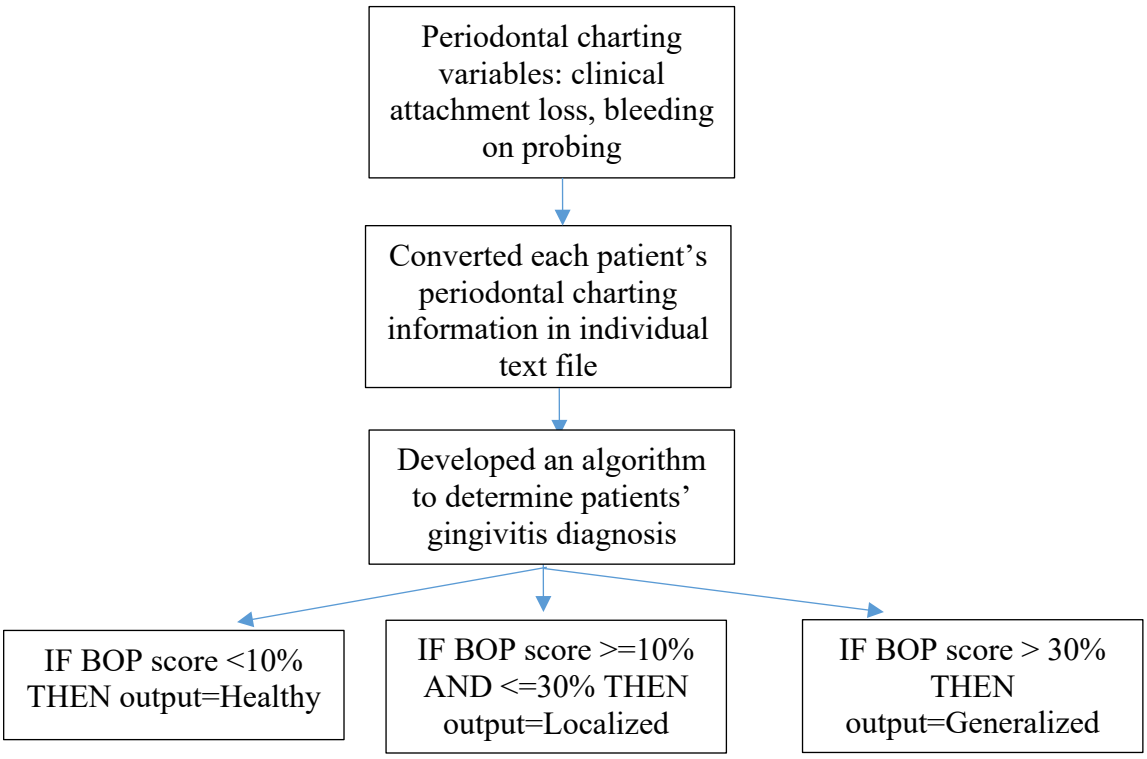


Figure 7: Illustration of encoding each patient in no gingivitis, localized or generalized gingivitis cases

3.7.3: Manual Review & Examining Performance of Algorithm

First, two clinical faculty members reviewed 50 common records and diagnosed patients' gingivitis status based on the BOP score. The inter-rater agreement between the faculties was 0.9 (Cohen's Kappa value) indicated excellent agreement. Next, each of the reviewers reviewed 150 records independently, which resulted in an overall dataset of 350 cases. Next, the diagnoses automatically generated by the algorithm were compared with the reviewers' diagnoses. Based on the computer algorithm's ability to correctly diagnose gingivitis cases, true positives, false positives, and false negatives were calculated. Using these measures, the gingivitis_diagnosis.py algorithm's precision, recall, and f-measure were calculated using the formulas described in Table 8.

Table 8: Evaluating the performance of the gingivitis_diagnosis.py algorithm on the testing dataset

Evaluation measures	Formulas
Precision	true positive / (true positives + false positives)
Recall	true positives / (true positive + false negatives)
F-measure	2 * (precision * recall) / (precision + recall)

3.8: Develop and Test an Algorithm to Diagnose Patients' Periodontitis Status from Periodontal Charting Findings.

3.8.1: Periodontitis Diagnosis Criteria

The case-definition developed by the AAP and CDC to classify a patient's periodontitis status into a mild, moderate, or severe periodontitis case was used as described in Chapter 2.2. Two periodontal charting findings that include 1) PPD (between the distance from the free gingival margin to the bottom of the sulcus or periodontal pocket), and 2) CAL to diagnose periodontitis were used [33, 126]. The periodontitis diagnostic criteria are described in the following paragraph (also described in Chapter 2.3).

- Severe periodontitis: Presence of two or more interproximal sites with ≥ 6 mm CAL (not on the same tooth) and one or more interproximal site(s) with ≥ 5 mm PPD.
- Moderate periodontitis: Presence of two or more interproximal sites with ≥ 4 mm clinical CAL (not on the same tooth) or two or more interproximal sites with PPD ≥ 5 mm, also not on the same tooth.
- Mild periodontitis: Presence of two or more interproximal sites with ≥ 3 mm CAL and two or more interproximal sites with ≥ 4 mm PPD (not on the same tooth) or 1 site with ≥ 5 mm.

Detailed description on the development of case definitions for periodontitis is explained in Chapter 2.3.

Table 9: Case-definition of periodontitis used in epidemiological studies to estimate prevalence of periodontitis

Severity (periodontitis)	Clinical Attachment Loss (CAL)	AND/OR	Periodontal Pocket Depth (PPD)
Mild	≥ 2 sites with ≥ 3 mm CAL (not on the same tooth)	AND	≥ 2 sites with ≥ 4 mm PPD (not on the same tooth)
	OR		
	≥ 2 sites with ≥ 3 mm (not on the same tooth)	AND	≥ 1 site with ≥ 5 mm PPD (not on the same tooth)
Moderate	≥ 2 sites with ≥ 4 mm (not on the same tooth)	OR	≥ 2 sites with ≥ 5 mm PPD (not on the same tooth)
Severe	≥ 2 sites with ≥ 6 mm (not on the same tooth)	AND	≥ 1 site with ≥ 5 mm PPD (not on the same tooth)

3.8.2: Developing the Algorithms

Three computer algorithms were developed (*Severe_Periodontitis_Classifier.py, Moderate_Periodontitis_Classifier.py, and Mild_periodontitis_Classifier.py*) to classify a patient's periodontitis status into no periodontitis, mild, moderate, or severe case. The detailed steps involved in developing these algorithms are described below.

3.8.2.1: Reading Text Files

First, the *"mkdir"* function was used to create a directory, and the *"os"* function to create two new folders 1) Severe_Cases and 2) Others. These two folders were created because the "*Severe_Periodontitis_Classifier.py"* read each text file based on the severe periodontitis criteria (see above: Criteria used to diagnose patients' periodontitis status) and transferred these files into either *"Severe_Cases"* or *"Other"* folder. Next, the *"Re.search()"* function was used to search only for files with *.TXT* extension (text files) in the folder. In addition, the "*open (files)"* function was utilized to save CAL and PPD information in a temporary variable, *'data'*.

3.8.2.2: Storing CAL & PPD Information in Temporary Variable

Next, "*IF"* and "*ELSE"* rules, together with "*FOR"* loop, were used to store a patient's CAL values to later determine whether this patient had severe periodontitis. *"IF"* and *"ELSE"* statements were used to find the word *"ATTACH"* in the stored text file. Once the *"ATTACH"* keyword was found, it performed the following function. If the algorithm did not find the *"ATTACH"* keyword in the current text file, then it would move on to the next file.

If *"ATTACH"* was present in the text file then the algorithm looked for the *"Current_Site_Name"* which represented the tooth number by the pre-defined index value. In each text file, a patient's tooth number was recorded in front of *"Site_Name",* and patients' six sites per tooth was recorded in front of "Value" (see Table 10). In this text file, *"Section: 0"* represents the *"buccal"* side of the tooth, and values 1, 2, and 3 represent three facial sites which include mesiobuccal, buccal (mid), and distobuccal respectively, followed by three lingual sites (mesiolingual, lingual (mid), distolingual). The 10^{th} index was used to determine the tooth number. As shown in Table 10, the tooth number "1" is recorded on the 10^{th} element. The algorithm saved this tooth number information in a temporary variable "*Current_SiteName"*. Next, once this tooth number was saved, it stored the maximum site value from those six sites (mesiobuccal, buccal, distobuccal, mesiolingual, lingual, distolingual) (see Table 10) in a temporary variable *"Sites_Affected_Attach"*. As shown in Table 8, the CAL values on six different sites are recorded on the *8^{th} index*. The algorithm collected the value of each of these sites and determined the maximum value. A rule that considered only one site (maximum CAL or PPD value) per tooth according to the AAP guidelines was also created [32, 33, 86]. This similar

process was repeated in order to find a patient's PPD information. This stored a patient's PPD information in a temporary variable *"sites_affected_pocket"*. By the end of this step, patients' CAL and PPD information are now stored for further analysis. The source code of this computer algorithm is described in Tables 47-49 (See Appendix section on pages 297-207).

Table 10: Indexing approach to locate tooth number and clinical attachment loss value in a periodontal charting text file

Patient's clinical attachment loss and periodontal pocket depth information recorded in a text file

SiteName: **1** →

Section: 0

Value 1: 3

Value 2: 3

Value 3: 3

SiteName: 1

Section: 1

Value 1: 4

Value 2: 2

Value 3: 3

SiteName: 2

Section: 0

Value 1: 2

Value 2: 2

S→0 element
i→ 1st element
t→ 2nd element
e→ 3rd element
N→ 4th element
a→ 5th element
m→ 6th element
e→ 7th element
: → 8th element
→ 9th element
1→10th element

→ 10th element is "1"

Value 3: 4

3.8.2.3: Criteria to Classify Patients' Periodontitis Status

From the aforementioned steps described, a patient's CAL and PPD information recorded on each site is stored in temporary variables. Next, the following rules were used to determine if a patient belonged to the "Severe" case. If the CAL and PPD information met the following criteria, then this text file was moved to the "Severe_Case" folder; and if not, then the file was moved to the "Other" folders. The rules utilized to perform this step are described in the Table 11.

Table 11: Criteria to determine whether the patient belonged to the severe or other periodontal disease categories

IF Sites_Affected_Attach >= 2 AND Sites_Affected_Pocket >= 1: THEN Severe Case ELSE: Others

After completing these steps, the algorithm placed severe cases in the *"Severe_case"* folder, and the rest of the text files in the *"Other"* folder. Next, similar algorithms were developed using the logic described in Table 12 to classify patients into moderate and mild groups. Patient text files that did not belong to severe, moderate, or mild cases and did not meet the criteria were considered as healthy cases. Figure 8 shows an illustration of classifying patients' periodontitis status into no periodontitis, mild, moderate, and severe periodontitis cases. The source code of this computer algorithm is described in Tables 47-49 (See Appendix section on pages 297-207).

Table 12: Criteria used to classify a patient's periodontal disease status into moderate, mild or healthy category

1)	IF Sites_Affected_Attach >= 2 or Sites_Affected_Pocket >= 2: THEN Moderate_Cases ELSE: Other
2)	IF (Sites_Affected_Attach >= 2 and Sites_Affected_Pocket >= 2) or THEN Moderate_Cases ELSE: Healthy

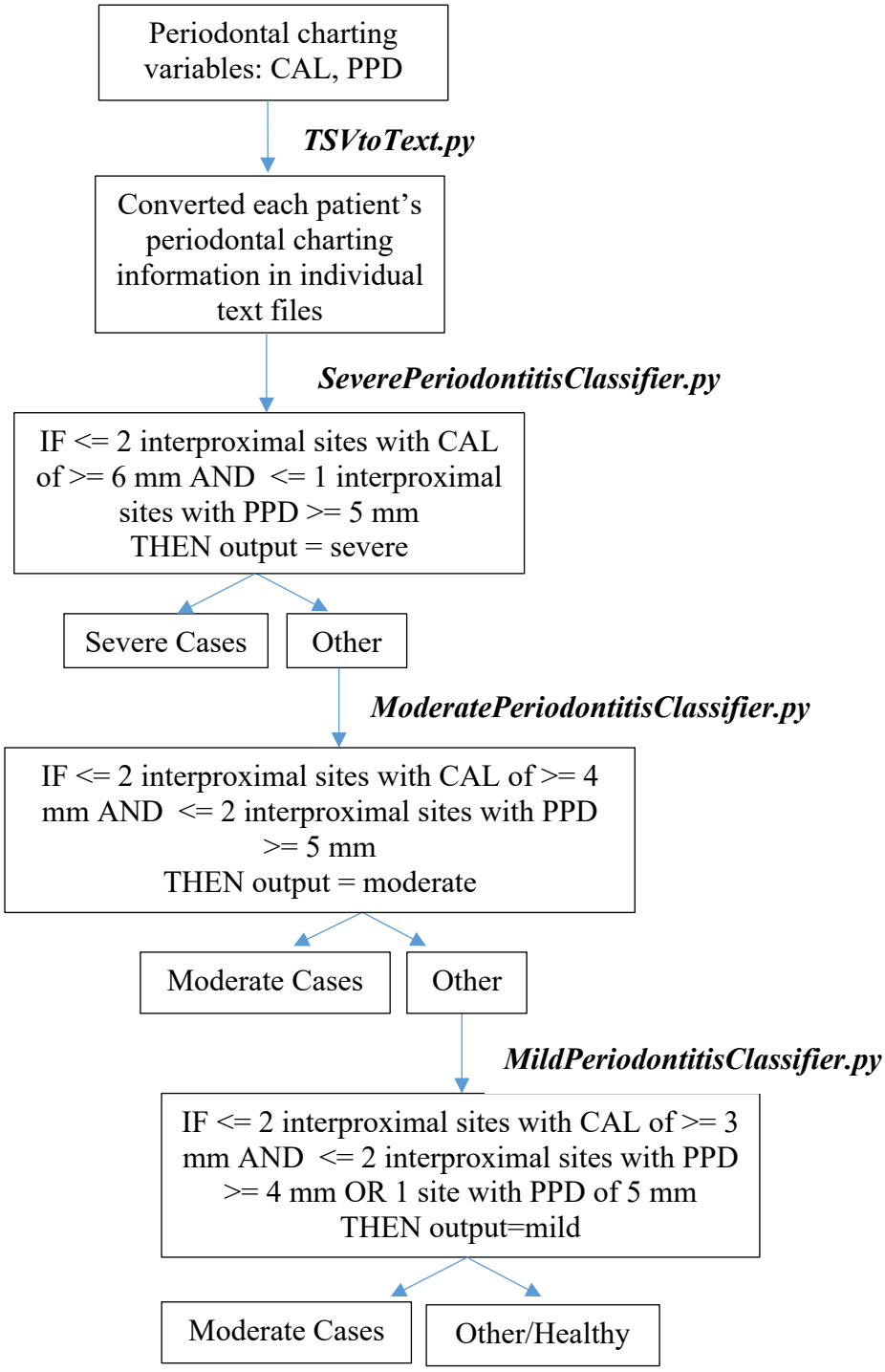


Figure 8: Illustration of determining a patient's periodontitis diagnosis into healthy, mild, moderate, and severe cases

3.8.3: Manual Review & Examining Performance of Algorithm

First, two clinical faculties reviewed 50 common records and diagnosed patients' periodontitis status based on the rules described in Tables 11, 12. The inter-rater agreement between the faculties was 0.9 (Cohen's Kappa value) which indicated excellent agreement. Next, each of the reviewers reviewed 150 records independently, which resulted in an overall dataset of 350 cases. Next, the diagnoses automatically generated by the algorithm were compared with the reviewers' diagnoses. Based on the computer algorithm's ability to correctly diagnose gingivitis cases, true positives, false positives, and false negatives were calculated. Using these measures, the *periodontitis_diagnosis.py* algorithm's precision, recall, and f-measure were calculated using the formulas described in Table 8.

3.9: Retrieving Clinician-Recorded Pd Diagnosis from Periodontal Evaluation Forms

Clinician-recorded PD diagnoses are stored in the free-text format within the periodontal evaluation forms. Extracting information out of the free-text data in an analyzable format is difficult due to the variation of clinical language used while documenting diagnosis. Therefore, a text-mining approach was developed (*Periodontal Disease Diagnosis Extractor.py*) to extract patients' PD diagnoses in a structured format. *Periodontal Diagnosis Extractor.py* consisted of several Python libraries such as the Natural Language Toolkit, Version 3.5 (NLTK), string-matching algorithm, regular expression, and pandas. The NLTK library was used to clean and pre-process the text data. First, the dataset was transformed into lower cases and tokenized. Tokenization is the process of splitting a text corpus into sentences that act as the first level of tokens that comprise the corpus. This process is also known as sentence segmentation because it was attempted to segment the text into meaningful sentences. Then the text was segmented by removing special characters and specific delimiters between sentences such as period (.), newline character (\n), and semi-colons (;).

Next, a total of 12 temporary variables (see Table 13) were created. As shown in Table 12, "PD_Disease" stored a patient's type of PD information (gingivitis or periodontitis), "PD_Severe" stored the severity of PD (mild, moderate, or severe), "PD_Location" stored the affected region (maxillary, mandibular, teeth number), "PD_extent" stored the extent of the disease (localized/generalized), and "PD_onset" stored the onset of the disease (acute/chronic)), Patient_ID, Patient_Birth, Patient _Race, Patient _Gender, and Patient _Insurance stored demographics information.

Table 13: Variables to save patients' periodontal disease (PD) severity, location, severity, extent and patient demographics

PD_Location, PD_Region, PD_Type, PD_Documentation_time, PD_Severity, Patient_ID, PD_diagnosis, Patient_DOB, Patient_race, Patient_Gender, Patient_insurance.

To extract a patient's detailed PD information based on the disease type, severity, extension, location, and region, the approximate string-matching algorithm (ASM) was used from the NLTK library. The ASM helps to find similar text or directory even though spelling or grammatical errors are present in the text. The approximate string search is formulated to find the text or dictionary of size *"N"* of all the words that start or match with the given word while considering all the *"K"* possible differential errors. This algorithm works on the "*Levenshtein distance"* concept. The "*Levenshtein distance"* is a metric to measure how apart two sequences of words are. Typically, in this logic, a user is asked to set a percentage of the match per requirement. For example, if the word *"periodontitis"* is written with spelling errors such as "perriodontitis" or "poridontitis" (see Table 15). In this example, there is an additional 'r' written in "periodontitis" which is 93% similar to the provided span of the word (periodontitis). Therefore, the ASM logic was able to detect these variations present in the clinical text and identify them successfully. Similarly, the ASM algorithm was utilized to automatically extract a patient's disease type, severity, location, and extent information. The span of texts used to perform ASM is described in Table 14.

Table 14: Words and span of text used in the Periodontal Disease Diagnosis Extractor.py algorithm to perform approximate string-matching function

Concept	Example of words and span of text
Disease type: gingivitis	gingivitis, inflammation of gingiva
Disease type: periodontitis	periodontitis, bone loss
Disease severity: mild, moderate, severe	mild, moderate, severe, mild to moderate, moderate to severe.
Disease extension: acute or chronic	acute, chronic

Table 15: Example of approximate string-matching algorithm using Levenshtein distance concept to find strings that match a pattern

Example clinical note: Patient is having a generalized horizontal bone loss. This patient is having moderate to severe perriodontitis. Span of text for string matching algorithm: "periodontitis" Insertion: "perriodontitis" Deletion: 'r' Substitution: None Match with the span of text=93% Output: disease type: periodontitis; severity: moderate

However, for many cases, one of these categories (disease type, disease severity, disease extension) was not recorded in the "diagnosis" section of the periodontal evaluation form. Therefore, to utilize EDR data to its optimum level, if any of the above described categories were not available, then the algorithm outputted "not specified" for the missing category. *Periodontal Disease Diagnosis Extractor.py* was able to classify patients' PD status from most superficial to detailed when the information is present. Figure 9 shows a breakdown of each classification category. The source code of this computer algorithm is described in Tables 50. (See Appendix section on pages 208-216). The performance of this

NLP algorithm was tested on a gold standard of 350 patients using the evaluation matrix as described in Table 8.

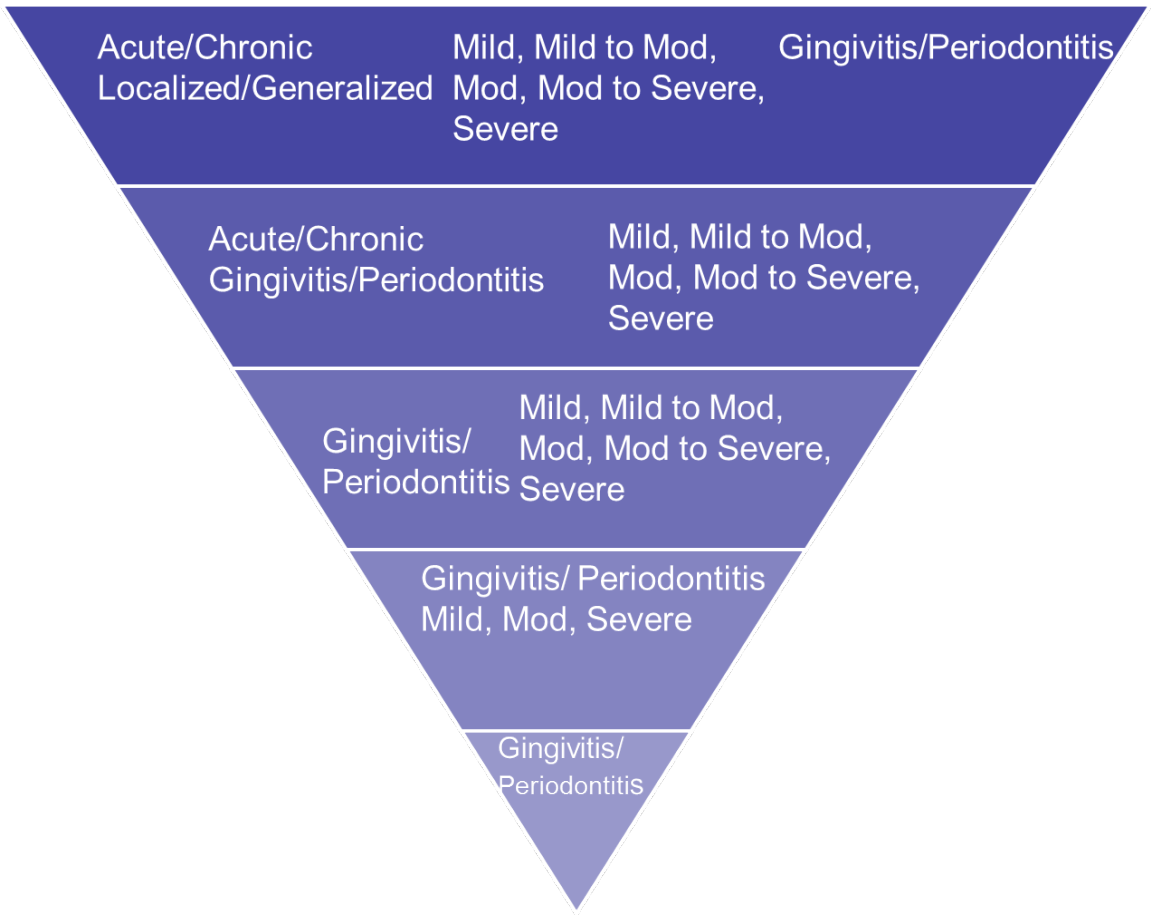


Figure 9: Bottom-Up approach to extract patients' clinician-recorded diagnoses based on disease type, disease severity, disease location, and disease extension

3.10: Extracting Patients' Periodontal Disease Diagnoses from Their First Comprehensive Oral Evaluation.

After performing the aforementioned steps, patients' PD diagnoses based on disease severity was classified automatically. As described in Chapter 4.6.2, patients' PD diagnoses were determined based on the date of their visits. Therefore, if a patient has more than one visit, then, he/she will have more than one diagnosis. For example, a patient (ID=00iar1) visited IUSD clinics on 11/1/2010, and 2/3/2013 and had two periodontal charting findings present. Then, this patient has two output files 1) 00iar1_11012010_Mild Periodontitis.txt, and 2) 00iar1_02032013_Moderate Periodontitis.txt. Patients' diagnoses from all of their visits were extracted in this fashion to track their diagnosis change over time. However, to report PD prevalence in the study population, only information from patients' first visits are required. Therefore, to obtain patients' PD diagnoses from their first completed COE, a computer algorithm *"First_COE_Information_Extractor.py"* was developed. The detailed steps taken to develop this algorithm are described in the following section. The source code of this computer algorithm is described in Table 52. (See Appendix section on page 220).

3.10.1: Importing Files

First, two Python libraries named Pandas and NumPy (support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays) were imported. Next, two files were imported: 1) .TSV file that has a patients' PD diagnosis, and date in MM/DD/YY format when patients' charting findings were recorded and 2) A treatment history subset in which patients' procedure codes (COE: D0150) and date of completed treatments were present (*Treatment_History_Subset.tsv*).

3.10.2: Creating a New Variable-offset Date

To extract patients' PD diagnoses from their first COE, the periodontal chart completion date with their COE completion date were compared. It was observed that, it takes on average approximately two to three appointments within a three months period to complete the COE (confirmed this by clinical faculties). Therefore, despite having a logic that compares the date of COE completion with the date of periodontal charting (exact date match), the variable "Offset_Date" was created. "Offset_Date" took a patient's COE completion date and saved this information in a variable 'ModifiedDateTime'. Next, by using (syntax: df_procedures ['Offset Date'] = df_procedures['ModifiedDateTime'].apply(lambda x: x-pd.DateOffset(months=3)), three months before and after the COE completion date, and 3 months after the COE completion date was saved in the "Offset_Date" variable.

3.10.3: Comparing "Offset_Date" with the periodontal charting completion date

To find patients' COE completion date information, two files were matched by the patient IDs. Patient IDs from the periodontal charting findings were stored in a variable *"patient_ID_perio"*. Then, the algorithm searched for this patient ID in the treatment history subset file. Once the algorithm found a similar patient ID, then it compared the charting finding date with the offset date. If the date of first COE matched with the date of documentation of charting, then the algorithm outputted as "TRUE", and if not, then outputted as "FALSE". The source code of this computer algorithm is described in Tables 51. (See Appendix section on pages 208-216).

3.11: Assessing the Completeness of Periodontal Disease Variables in the Electronic Dental Records

Completeness

Completeness of data is defined as the extent to which data is sufficient to answer a proposed research question [36, 133, 134]. Variables to answer this study's research question are demographic variables such as age, gender, race/ethnicity, insurance status, periodontal findings that contribute to PD diagnosis, and clinician-recorded PD diagnoses in the periodontal evaluation form. To examine completeness, the presence or absence of this information was determined in the EDR. The proportions of present values by the total number of patients who received COE between January 1, 2009, to December 31 2014 was calculated.

3.12: Assessing the Concordance Between Diagnosis Generated from Periodontal Findings and Clinician-Recorded Diagnoses.

The definition of concordance which was proposed by Weiskopf et al was used [134] . Data is considered concordant when there is an agreement or compatibility between the same information present in two data fields. This may mean that two elements recording the same information for a single patient have the same value. Therefore, to determine concordance between diagnoses generated from periodontal findings and clinician-recorded diagnoses, the percentage agreement was calculated. The percent agreement was calculated for those patients who had both diagnoses generated from charting and clinician-recorded diagnoses available. There were a total of 10,406 patients whose both charting and clinician-recorded diagnoses were available. Before comparison, the study confirmed that both diagnoses generated from findings and clinician-recorded diagnoses were recorded during the same time period.

Developing a gold standard of 125 patient records

Next, to determine the reasons for the disagreements and agreements, a reference standard dataset was developed consisting of 125 manually reviewed diagnoses. It was developed by conducting four rounds of the manual review process. During the first round, ten records were selected that were independently reviewed by two clinical faculties (Dr. Dan Shin and Dr. Lisa Willis) at the IUSD. They reviewed patients' detailed radiographic findings, gingival patterns, signs and symptoms information documented in periodontal evaluation forms. The detailed description of manual review guidelines is described in the Appendix section (see page 221). Next, the inter-rater agreement (percent agreement) was performed between diagnoses determined by both of the reviewers. For the first round of

the review, 50% agreement was observed. Then the disagreed diagnoses were discussed by both reviewers and a consensus was made. During the second and third rounds, both of the reviewers reviewed twenty more records independently and the agreement was calculated. A 95% agreement was achieved. In summary, both of the reviewers reviewed the same 30 records and the agreement between diagnoses determined by reviewers was excellent (95%). During the fourth round, each reviewer reviewed 50 records independently which were not the same patients. Among these 50 records, five records were the same records to ensure excellent agreement between reviewers. The end of this step resulted in a reference standard of the total 125 patient records. The breakdown of these 125 records is described in Figure 10. Last, the percent agreement between the diagnosis recorded in the gold standard dataset versus clinical record diagnosis was determined.

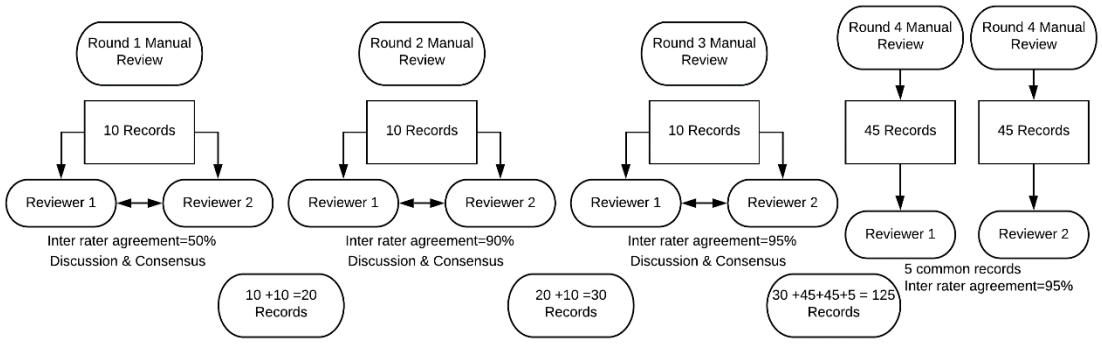


Figure 10: Manual review process to determine agreement between diagnoses generated from findings and clinician-recorded diagnoses

3.13: Data Analysis

The final dataset included patients who received COE between January 1 2009, and December 31, 2014 at IUSD. Other information that fall outside of this time period was also included. from Therefore, the observation time of this study was from June 1, 2005 to August 1, 2019.

First, age was stratified by 18 to 29 years, 30 to 44 years, 45 to 64 years, and 65 years and older. Patients' gender was classified as female, male, transgender, other, and unknown. Race/ethnicity was classified as Caucasian, African American, Hispanic or Latino, Asian, Multiracial, American Indian, Pacific Islander, unknown, and missing race (section left blank). Patents' insurance status was classified into four categories; self-pay, private insurance, public insurance, and unknown.

Next, descriptive statistics was performed with 95% confidential intervals on PD diagnosis, age, gender, race/ethnicity, insurance status, and the number of procedures received per year from August 1, 2005 to August 1, 2019. Next, the completeness of each data variable was calculated using proportions. the total number of patients who received COE from 2009 to 2014 as a denominator value was used to measure completeness. Then concordance was measured between PD diagnoses generated through periodontal findings and clinician recorded diagnoses. To measure concordance, percent agreement was calculated.

The PD prevalence was calculated using clinician-recorded diagnoses. The prevalence of PD was calculated by age group, gender, insurance status, and race/ethnicity. The difference between each category was examined using the Chi-Square test. The

associations between PD (gingivitis, and periodontitis) and demographic information were examined using the chi-square statistical test. The p value was set at 0.05.

4: Aim 2 & Methods

Aim 2

Determine the feasibility of tracking the change in a patient's periodontal disease diagnosis using electronic dental record data.

Hypothesis

It is hypothesized that EDR data will enable us to track the change in periodontal status for at least 50% of patients who underwent COE. The rationale is that at least 50% of patients will have an average of two visits during the six-year study period.

To test this hypothesis, first, average visits for patients who received COE between January 1, 2009 to December 31, 2014 were generated. Next, a computer algorithm was developed that generated the number of patients whose disease status did not change over time, and patients whose disease status regressed or progressed over time. The study results will help in determining the potential of EDR data to predict a patient's disease progression over time and determine PD treatment outcomes.

4.1: Data Source & Study Design

Data Source

EDR (axiUm®-EXAN) data from the IUSD pre-doctoral clinics to conduct this study. The EDR system was implemented at IUSD in the year 2005 was utilized. Patients who visited IUSD and IUSD satellite clinics and received at least one comprehensive oral evaluation (COE) between January 1, 2009, to December 31, 2014 were retrieved. The patients' visit information that may fall outside this time period were also included in this study. For example, if a patient received COE in 2010 and received treatments in 2007 and 2015 then, information from 2007 and 2015 would also be included in this study. This time period was selected because it is aimed to compare prevalence generated in a recently conducted epidemiological study [33]. This study has estimated the prevalence of periodontitis in the United States between 2009 and 2014. Selecting this similar timeline would allow an effective comparison between this study results and with the results of epidemiological study [33].

Study Design

The design of this study was a retrospective cohort study.

4.2: Inclusion & Exclusion Criteria

Inclusion criteria

This study dataset included patients who underwent COE between January 1, 2009, to December 31, 2014, and who were 18 years or older during the time of their first completed COE procedure. Four main data tables in axiUm® were utilized: 1) patient demographics (date of birth, gender, race, insurance), 2) procedure codes (treatments provided), 3) periodontal charting (CAL, PPD, BOP), and 4) periodontal evaluation form (clinician-recorded PD).

Exclusion criteria

Patients who were less than 18 years old and who did not receive COE during the study time period (January 1, 2009, to December 31, 2014) were excluded.

4.3: Study Variables

Demographics

Patients' demographic information such as date of birth, gender, and race/ethnicity, was retrieved. Patients' age was calculated from their date of birth information recorded during their first COE examination.

Periodontal evaluation form

From the periodontal evaluation forms, patients' clinician-recorded diagnoses were retrieved. PD diagnoses were extracted from patients' every visit. The clinician-recorded diagnosis extractor computer algorithm developed in Aim 1 (see Chapter 4.10) was utilized to extract PD diagnoses in a structured format.

4.4: Developing & Testing Periodontal Disease Change Counter.py Algorithm

To determine PD change over time, patients' PD diagnoses were sorted (obtained from Aim 1-Chapter 4.2) by their visit dates. Second, an algorithm was written in Python, *"Diagnosis change overtime classifier.py"* as described in the following paragraph to automatically classify a patient's disease progression status into one of the following categories.

1) disease status did not change between patients' first and last visits,

2) disease progressed between patients' first and last visits, and

3) disease improved between patients' first and last visits.

"Diagnosis change overtime classifier.py" algorithm consisted of several Python libraries that include "NLTK", "string", "regular expression", and "pandas". By using these libraries, first the algorithm read the text file using *read* function and saved "disease type (gingivitis or periodontitis)" "severity (mild, mild to moderate, moderate, moderate to severe, severe)", and "disease extension (localized/generalized)" in temporary variables *Severity_List* and *Disease_Type,* respectively. Next, the algorithm created two temporary variables *"From"* and *"To"*. The algorithm determined the difference of date between the two visit dates. If these two dates were different (differences have to be 90 days apart), then, one of the PD diagnoses recorded during the first date was placed in the *"From"* temporary variable. The diagnosis recorded at the latest date was placed in the *"To"* temporary variable. Next, the algorithm determined if these two dates recorded in the *"From"* and *"To"* variables were similar or not. If they were similar, then the algorithm skipped these records and went to the next available date. If there was no other diagnosis present, then it went to the next row (patient ID). The algorithm determined the disease

change status by using a regular expression function. If the dates stored in *"from"* and *"to"* were different, then the program determined the *"disease type"*, *"severity"*, and *"disease extension"* in the variables *"from"* and *"to"*.

If the *"disease type"*, *"severity"*, and *"disease extension"* were similar, then this information was saved in a variable *"disease_change_status".* Next, the algorithm looked for the disease information stored in *"disease_change_status"* for all the patients and patients who had similar changes in the disease and counted them using the "*for loop"* function. For example, if there were 200 patients whose disease status was mild generalized periodontitis during their first and second visits, then the algorithm would output the result as *"from-mild generalized periodontitis" to "mild generalized periodontitis"* = 200. Similarly, if 300 patients' disease status changed from mild generalized periodontitis to moderate generalized periodontitis, then the number of patients with a similar change in disease status would be output as *"from-mild generalized periodontitis" to "moderate generalized periodontitis"* = 300. The algorithm also considered *"no information present"* fields as a separate group. For example, if 200 patients' diagnoses did not have a mention of *severity region (localized/generalized)* and mentioned mild periodontitis then the algorithm would output as *"from-mild periodontitis" to "moderate* generalized periodontitis" = 200. The source code of this computer algorithm is described in Tables 51, 52. (See Appendix section on pages 216-220).

Next, each patient was classified into one of the three categories: 1) patients' PD status did not change over time, 2) patients' PD progressed over time, and 3) patients' disease status improved over time (see Table 15). For instance, if the patient has transitioned from mild to moderate gingivitis, then that patient is categorized as belonging

to “patients PD progressed over time”. Detailed descriptions of each group are provided in Table 16. This led to a total of 558 categories.

Table 16: Example longitudinal patient data representing change in periodontal disease between two subsequent dental visits

Patient IDs	Visiting date 1	Diagnosis 1	Visiting date 2	Diagnosis 2
1	1/5/2010	Chronic mild periodontitis	3/12/2012	Chronic moderate periodontitis

4.5: Manual Review & Examining Performance of Algorithm

A gold standard of 125 randomly selected patient cases was created to evaluate the performance of *"Diagnosis change overtime classifier.py*. These patients' clinician-recorded diagnoses by their visit dates were manually reviewed to determine the disease change to classified them into one of the three categories: 1) patients whose disease status did not change, 2) patients whose disease status progressed, and 3) patients whose disease status improved. Next, the patients' PD status change counts generated by the *"Diagnosis change overtime classifier.py"* algorithm was compared against the gold standard. Based on the computer algorithm's ability to correctly generate counts, true positives, false positives, and false negatives were calculated. By using these measures, the algorithm's precision, recall, and f-measure were calculated using the formulas described in Table 8.

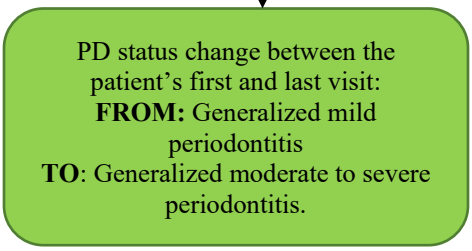
4.6: Data Analysis

Descriptive statistics with 95% confidential intervals was performed on the number of periodontal charting, and clinician-recorded diagnoses documented between June 1 2005, to August 1, 2019 for the patients who received at least one COE between January 1, 2009, and December 31, 2014. The average days, months, and years between patients' first, and second; first and third; and first and fourth visits were calculated. This test helped us identifying how frequently patients' clinician-recorded diagnoses were available to determine their disease change over time. The frequency count and the number of patients by the observation time between their first and last visits were generated. The frequency counts were generated in the following six categories: 1) no-follow-ups, 2) up to 5 years, 3) >5 and <=10 years, 4) >10 and <=15 years, 5)>15 and <=20 years, and 6) more than 20 years.

Next, the frequency count of the number of patients whose disease status did not change, disease status progressed, and disease status improved from their first to the last visit using patients' clinician-recorded diagnoses were also generated. As shown in Table 18, one of the patients had a total of 7 clinician-recorded diagnoses between June 1 2005, to August 1, 2019. As shown in Table 18, the orange color errors show the disease change after every consecutive visit and the green arrow shows the patient's disease status change from his first and last visit. For every patient who had at least two clinician-recorded diagnoses, disease status change between their first and last visits that fall under June 1 2005, to August 1, 2019 time period were determined.

Table 17: Generating counts of patients whose periodontal disease status did not change, progressed or improved from their first to their last visit

Visit 1: generalized mild periodontitis	Visit 2: localized mild periodontitis
Visit 2: localized mild periodontitis	Visit 3: localized mild periodontitis
Visit 3: localized mild periodontitis	Visit 4: localized moderate periodontitis
Visit 4: localized moderate periodontitis	Visit 5: generalized moderate periodontitis
Visit 5: generalized moderate periodontitis	Visit 6: localized moderate to severe periodontitis
Visit 6: localized moderate to severe periodontitis	Visit 7: generalized moderate to severe periodontitis



5: Results

5.1: Number of Treatments Received by Patients

Total number of unique patients

This study cohort consisted of 28,908 unique patients who received at least one COE between January 1, 2009, and December 31, 2014. Out of the 28, 908 patients, 22,191 (77%) patients received care at the IUSD predoctoral comprehensive care clinics. The remaining 6,747 (23%) patients received care at the IUSD satellite clinics which include Regenstrief (2,719 (9%)) Cottage Corner (2, 020 patients (7%)), Grassy Creek (1, 169(4%)), and University hospital (809(3%)) dental clinics (See Table 19).

Table 18: Number of patients that visited IUSD clinics and IUSD satellite clinics between January 1, 2009 to December 31, 2014

IUSD clinics		22,191 (77%)
IUSD satellite clinics	Regenstrief	2,719 (9%)
	Cottage Corner	2,020 (7%)
	Grassy Creek	1,169 (4%)
	University Hospital	809 (3%)
	Total patients (IUSD satellite clinics)	6,717 (23%)
Total		28, 908 (100%)

Number of patients receiving comprehensive oral evaluations, periodic oral evaluations, periodontal maintenances, and periodontal re-evaluations between January 1, 2009 to December 31, 2014

Figures 11-16 display the number of COEs, POEs, PMs, and PREs received by IUSD patients between January 1, 2009 to December 31, 2014. These figures also show procedures performed outside of this time period (January 1, 2009 to December 31, 2014) because some patients who received these treatments during this study time period (January

1, 2009 to December 31, 2014) have also received treatments before or after this time period (anytime between June 1, 2005 and August 1, 2019). This study found that the maximum number of COEs (5,766) were received in 2011, POEs (2,897) in 2014, PREs (943) in 2012, and PMs (1,807) in 2013. These figures show the total number of multiple treatments received by 28,908 unique patients. The Appendix section (page 233) contains the first COE, POE, PRE, and PM (one treatment per patient) received by 28, 908 patients between January 1, 2009 and December 31, 2014.

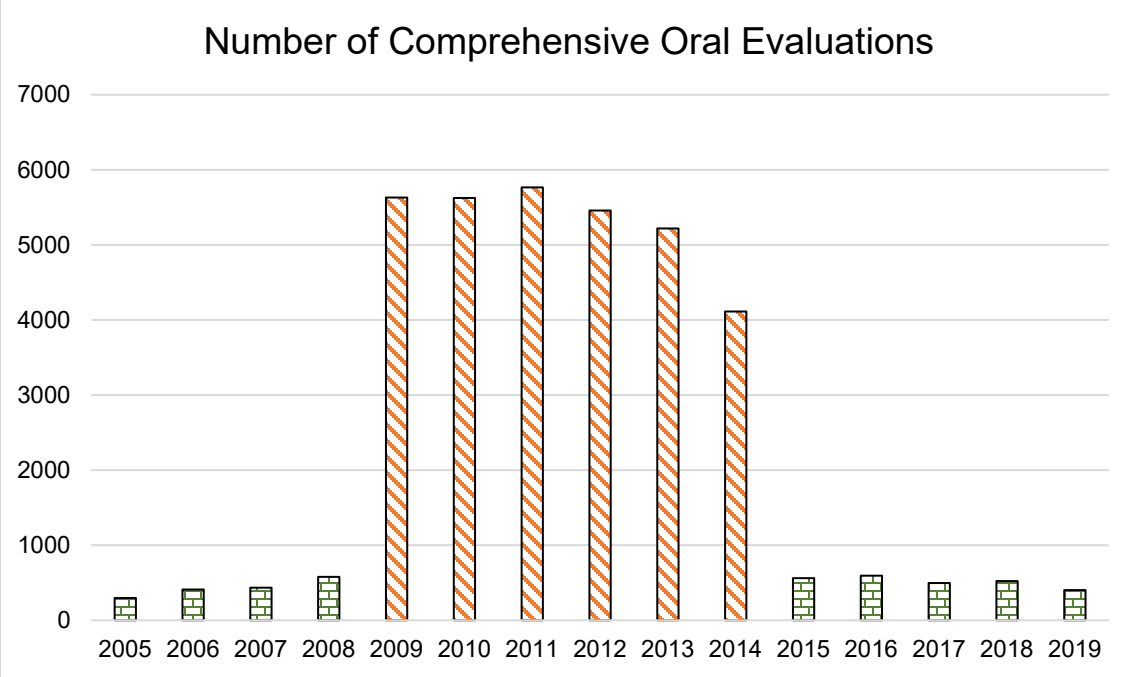


Figure 11: Number of comprehensive oral evaluations received at the Indiana University School of Dentistry clinics by 28,908 patients

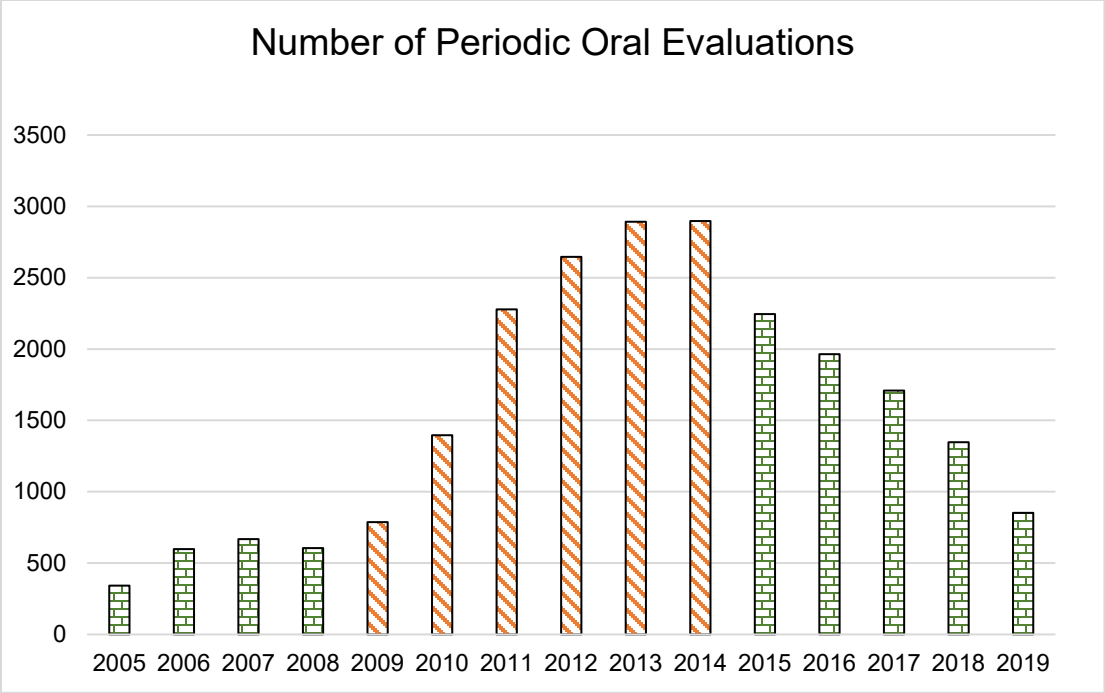


Figure 12: Number of periodic oral evaluations received at the Indiana University School of Dentistry clinics by 28,908 patients

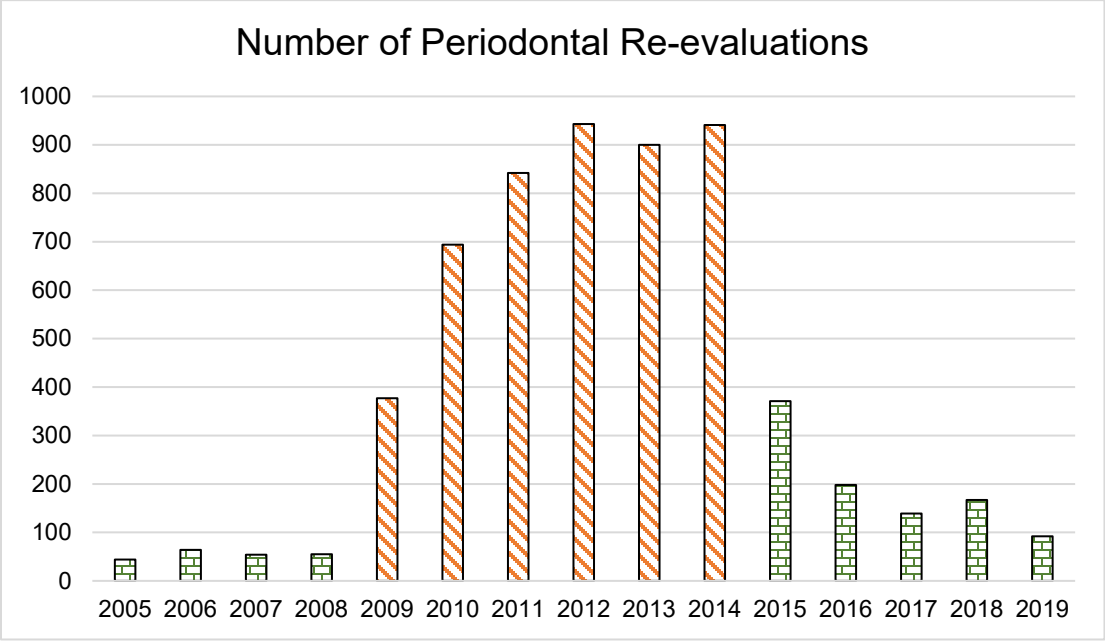


Figure 13: Number of periodontal re-evaluations received at the Indiana University School of Dentistry clinics by 28,908 patients

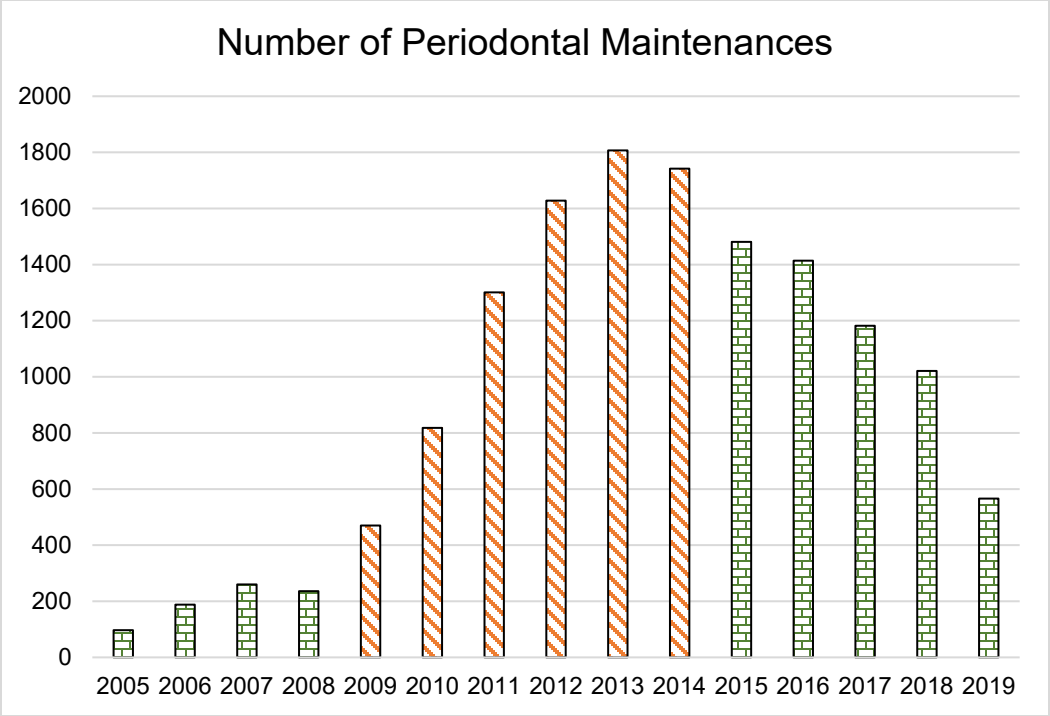


Figure 14: Number of periodontal maintenances received at the Indiana University School of Dentistry clinics by 28,908 patients

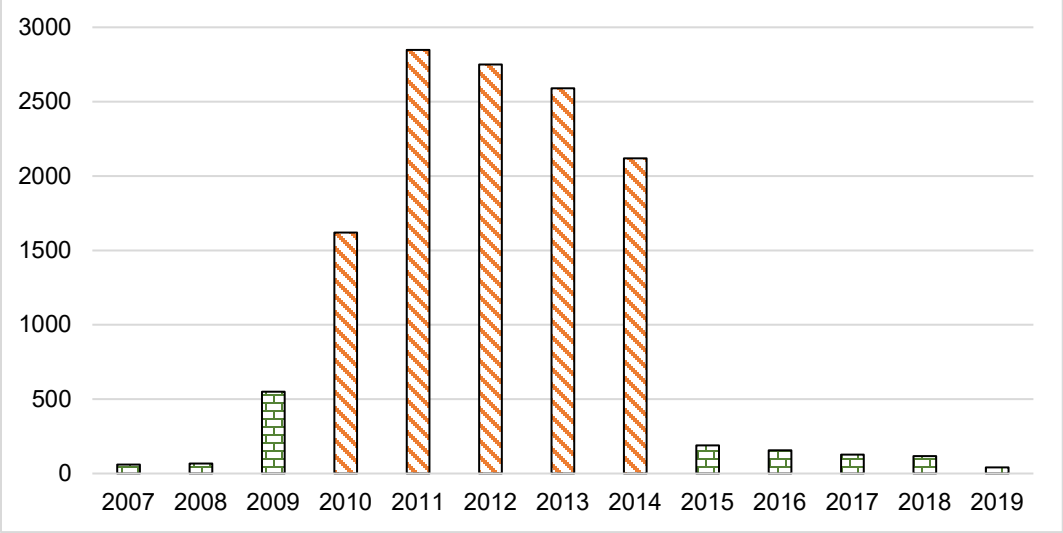


Figure 15: Number of periodontal evaluation forms received at the Indiana University School of Dentistry clinics by 28,908 patients

Figure 16 shows the number of periodontal evaluation forms and PD diagnoses documented between January 1, 2009, and December 31, 2014. It shows the comparison of the number of COE received versus documentation of PD diagnoses of these patients. The documentation of patients' PD diagnoses improved significantly over the period of time. PD diagnoses were least available (11%) for patients who received COE in the year 2009 and most available (66%) for patients who received COE in the year 2014. More than half (54%) of PD diagnoses were missing for patients who received COE between January 1, 2009, and December 31, 2014.

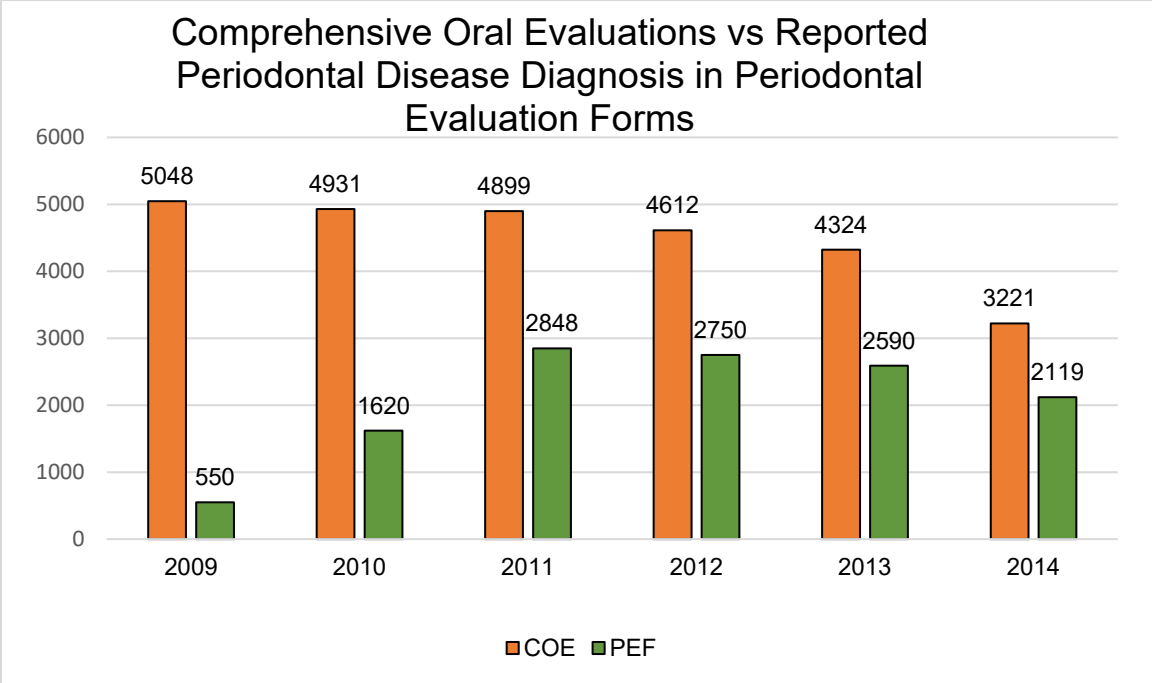


Figure 16: The number of comprehensive oral evaluations versus recorded periodontal disease diagnoses by unique patients in the periodontal evaluation forms

5.2 Data Quality Measure: Completeness

The completeness of the date of birth, gender, race, periodontal charting findings, treatment histories and clinician-recorded diagnoses using two denominators were calculated. One denominator used was 23,121 which is the total number of patients who received care only at IUSD. The second denominator used was 28,908, the total number of patients who received care at IUSD and its satellite clinics. When considering the denominator of 28, 908, patients' date of birth, and gender were documented for all 28,908 patients (100% complete). Race/ethnicity information was reported for 20,880 (71%) patients. Periodontal charting variables including BOP, CAL, and PPD were recorded for 22,880 (80%) of patients. Clinician-recorded PD diagnoses were recorded for 13,219 (46%) patients (see Table 21). When considering the total number of patients who received care at IUSD (not satellite clinics) as a denominator, the completeness improved for race (from 72% to 82%), periodontal charting (from 80% to 89%), and clinician-recorded diagnoses (from 46% to 53%) variables (See Table 20).

Table 19: Completeness of patient demographics, insurance, periodontal charting, and periodontal disease diagnosis who received care at Indiana University School of Dentistry clinics

Variables	Number of unique patients N= 23,122 (%)
Total Number of Patients	23,121 (100)
Date of Birth	23,121 (100)
Gender	23,121 (100)
Insurance	23,119 (100)
Race	18,907 (82)
Periodontal charting data (CAL, BOP, PPD)*	20,571 (89)
Clinician-documented PD Diagnosis	13,219 (53)

CAL = Clinical Attachment Loss, BOP= Bleeding on Probing, PPD = Periodontal Pocket Depth.

Table 20: Completeness of patient demographics, insurance, periodontal charting, and periodontal disease diagnosis who received care at Indiana University School of Dentistry clinics and satellite clinics

Variables	Number of unique patients (%)
Total Number of Patients	28,908 (100)
Date of Birth	28,908 (100)
Gender	28,908 (100)
Insurance	28,905 (100)
Race	20,628 (71)
Periodontal charting data (CAL, BOP, PPD)	22,880 (80)
Clinician-documented PD Diagnosis	13,219 (46)

5.3: Patient Demographics and Characteristics

Table 22 describes the age distribution of the patient population. The mean age of the patient population was 46 years old (standard error=0.09, standard deviation=16.74). The patient population consisted of more female patients (54%) than male (46%) patients (see Table 23). The majority of the patient population were Caucasians (49%) (see Table 24). As demonstrated in Table 25, 44% of patients had private insurance, 11% had government insurance, and 45% self-paid for the treatments.

Table 21: Age distribution of patients who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

Age Range (years)	N	(%)
18-26	4,727	(16)
27-36	5,310	(18)
37-46	4,745	(16)
47-56	6,055	(21)
57-66	4,597	(16)
67-76	2,430	(8)
77-86	914	(3)
87 & more	130	(0)
Total	28,908	(100)

Table 22: Distribution of patients Gender who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

Gender	N	(%)
Female	15,572	(54)
Male	13,242	(46)
Transgender	8	(0)
Other	86	(0.2)
Total	28,908	(100)

Table 23: Distribution of patients' race who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

Race/Ethnicity	N	(%)
Caucasian	14,038	(49)
African American	3,856	(13)
Hispanic	1,902	(7)
Asian	580	(2)
Other	164	(1)
Multiracial	41	(0)
American Indian	14	(0)
Pacific Islander	2	(0)
Unknown	31	(0)
Missing	8,280	(29)
Total	28,908	(100)

Table 24: Insurance information of patients' who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

Insurance	N (%)
Self-pay	13, 067 (45)
Private insurance	12, 751 (44)
Government insurance	3,090 (11)
Total	28,908 (100)

5.4: Patients' Periodontal Disease Diagnoses Determined from Periodontal Charting Findings

The patients' PD diagnoses using the case definitions of gingivitis and periodontitis were determined. First, the patients' gingivitis status using the BOP score without considering their periodontitis status was obtained. Similarly, patients' periodontitis status using CAL and PPD information without considering their gingivitis status was determined. As described in Table 26, 4,348 cases (62%) with no gingivitis, 5,621 cases (24%) with localized gingivitis, 2,909 cases (24%) with generalized gingivitis, 2 cases as unknowns out of 22,880 (100%) total patients were found. The study also found that 3,708 (16%) patients had no periodontitis, 182 (0.78%) had mild, 12,635 had moderate (55%), and 6,317 (27%) had severe periodontitis (see Table 27).

Next, patients' gingivitis and periodontitis diagnoses were considered together. If a patient had both gingivitis and periodontitis, then periodontitis diagnosis was considered first because it is a more severe condition than gingivitis. This study found 3, 194 (14%) patients as healthy, 438 (2%) as localized gingivitis, 76 (0.3%) as generalized gingivitis, 182 (0.8%) as mild periodontitis, 12,635 (55%) as moderate periodontitis, 6,317 (28%) as severe periodontitis, and 38 (0.2%) as unknowns out of 22,880 (100%) total patients (see Table 28).

Table 25: Patients' gingivitis status determined by calculating the bleeding on probing score from periodontal charting findings

Gingivitis	N	(%)
Healthy	14,348	(62)
Localized	5,621	(24)
Generalized	2,909	(13)
Unknown	2	(0)
Total (available data)	22,880	(80)
Missing	6,028	(20)
Total	28,908	(100)

Table 26: Patients' periodontitis status determined by calculating clinical attachment loss and periodontal pocket depth information from periodontal charting findings

Periodontitis	N	(%)
Healthy	3,708	(16)
Mild	182	(0.78)
Moderate	12,635	(55)
Severe	6,317	(27)
Unknown	38	(0)
Total (available data)	22,880	(100)
Missing data	6,028	(20)
Total	28,908	(100)

Table 27: Patients PD diagnoses after combining their gingivitis and periodontitis status

Gingivitis & Periodontitis	N	(%)
Healthy	3,194	(14)
Localized gingivitis	438	(2)
Generalized gingivitis	76	(0.3)
Mild periodontitis	182	(0.8)
Moderate periodontitis	12,635	(55)
Severe periodontitis	6,317	(28)
Unknown	38	(0.2)
Total (available data)	22,880	(100)
Missing data	6,028	(20)
Total	28,908	(100)

5.5: Evaluation of Algorithm's Performance

While comparing patients' PD diagnoses generated from the computer algorithms (*Subset Extractor.py*, *Gingivitis Diagnoser.py, Periodontitis Diagnoser.py*) against the gold standard, a 100% precision recall and f- measure was achieved. All records belonged to the true positive case which resulted in excellent accuracy. The reasons behind the excellent performance is that the data is structured and information in each text file is written consistently. For example, as described in Chapter 4.7.2, the indexing method was used to locate a tooth number in the text file. In each text file, each tooth was present on the 10^{th} element. As a result, computer algorithms were able to correctly identify all patient cases.

5.6: Clinician-Recorded Periodontal Disease Diagnoses

The clinician-documented diagnoses were available for a total of 13,219 patients (46%). Among these patients, 3,193 patients (24%) were diagnosed with mild gingivitis, 1,607 (12%) with moderate gingivitis, and 143 (1%) with severe gingivitis out of 13,219 available periodontal evaluation forms. Eighteen percent of patients (2,430) were diagnosed with mild periodontitis, 1,899 (14%) with moderate periodontitis, and 554 (4%) with severe periodontitis cases (see Table 28). Clinicians also documented patients' PD diagnoses with two additional severity categories such as "mild to moderate", and "moderate to severe". There were 247 (2%) patients with mild to moderate gingivitis, 62 (0.5) with moderate to severe gingivitis, 569 (4%) with mild to moderate periodontitis and 350 (3%) with moderate to severe periodontitis. There were 1,613 cases who were classified into only gingivitis, and 258 patients were classified into only periodontitis because of the lack of availability of these patients' disease severity information. There were 294 (2%) cases where the computer algorithm missed correctly identifying patients' either severity or disease status (see Table 29).

Table 28: Clinician-recorded patients' gingivitis and periodontitis diagnosis determined from the "diagnosis" section of the periodontal evaluation forms

Gingivitis/Periodontitis	N	(%)
Mild gingivitis	3,193	(24)
Mild to moderate gingivitis	247	(2)
Moderate gingivitis	1,607	(12)
Moderate to severe gingivitis	62	(0.5)
Gingivitis	1,613	(12)
Severe gingivitis	143	(1)
Mild periodontitis	2,430	(18)
Mild to moderate periodontitis	569	(4)
Moderate periodontitis	1,899	(14)
Moderate to severe periodontitis	350	(3)
Periodontitis	258	(2)
Severe periodontitis	554	(4)
Missing/no disease mentioned/algorithm error	294	(2)
Total (available data)	13,219	(100)
Missing data	15,689	(54)
Total	28,908	(100)

5.7: Manual Review & Examining Performance of Periodontal Disease Diagnoser.Py Algorithm

Table 29 shows the performance of the *Periodontal Disease Diagnoser.py* algorithm on correctly identifying disease type, disease concept, disease severity, disease region and disease extent from the clinical notes. A 99% precision for extracting disease type (gingivitis or periodontitis), and 97% precision of extracting severity information (mild, mild to moderate, moderate, moderate to severe, severe) was found.

Table 29: Performance of the Periodontal Disease diagnosis extractor for encoding disease type, extent, severity, and region

Concepts	Precision (%)	Recall (%)	F-measure (%)
Disease onset (acute/chronic)	100	99	99
Disease status	99	99	99
Disease extent (localized/generalized)	99	99	99
Disease region	97	97	97
Disease severity	97	96	96
Overall	**98**	**98**	**98**

5.8: Concordance Between Diagnoses Generated from Periodontal Charting Findings and Clinician-Documented Diagnosis

Patients' PD diagnoses (gingivitis/periodontitis) determined from periodontal charting findings with the clinician-recorded diagnoses were compared. Disease severity (mild, moderate, severe) category was excluded from this comparison because the severity level described in the AAP classification, and severity level used by clinicians are different. For example, AAP classified patients' PD status into mild, moderate, and severe, while clinicians also used two additional categories "mild to moderate" and "moderate to severe". The study found a 47% agreement (percentage agreement) between PD diagnoses determined from periodontal findings and clinician-recorded diagnoses (10,406 patients had both charting and clinician-recorded diagnoses available).

Agreement between gold standard dataset and diagnoses determined from findings, clinical-recorded diagnosis

When the gold standard dataset was compared with the clinician-recorded diagnoses, 81% agreement was achieved (see Chapter 4.12). While comparing diagnoses generated from periodontal findings with gold standard dataset diagnoses, 40% agreement was discovered.

Manual chart review to determine reasons for disagreement between diagnosis generated from findings and clinician-recorded diagnosis

As described in Chapter 4.12, two dentists (Dr. Dan Shin and Dr. Lisa Willis) manually reviewed 125 patients' periodontal evaluation forms to determine the reasons for the disagreement between diagnoses generated from findings and clinician-recorded

diagnoses. Reasons for disagreement found through the manual review process are described as following.

The cut-off criteria of the BOP score missed diagnosing 31% of gingivitis cases

While manually reviewing the disagreed diagnoses, it was observed that for some cases clinicians diagnosed patients in gingivitis category based on the intraoral findings (gingival color, contour), however, the BOP score was not high enough for these patients to diagnose them in the gingivitis category. This resulted in a disagreement which accounted for 31% (42 out of 125 cases) of the total disagreements. During the manual review process, clinicians were asked to document the reasons for their diagnoses. In these cases, they documented that the patients' clinical notes clearly stated patients having generalize inflammation, erythema, edema, and loss of stippling signs that confirms the gingivitis case. However, when the BOP score was calculated, the score wasn't high enough to diagnose these patients' in gingivitis case based on the case definition criteria [126].

Forty nine percent of gingivitis cases were diagnosed as periodontitis due to higher CAL and PPD value in patients' periodontal charting but did not have bone loss

For periodontitis, 71% of cases were found in which patients' periodontitis diagnosis was not accurately determined by the algorithm from the findings. It was observed that either these patients did not have periodontitis or had a milder form where the case definition classified these patients in more severe form than an actual severity (diagnosed by clinicians) of periodontitis. In the gold standard dataset, this study also found that 42 out of 125 records (34%) had solely gingivitis and not periodontitis. However, these

patients were diagnosed as periodontitis through the Periodontitis_Diagnoser.py algorithm based on the case definition criteria which resulted in disagreements.

Example comment from a manual reviewer:

> "Localized gingivitis- BOP is 10%, positive generalized visual evidence of inflammation, no evidence of radiographic bone loss, CAL is recorded, however not TRUE CAL attributable to periodontitis and may be false positives due to the limitations of axium. An argument can be made this is generalized gingivitis based on the student's description, however from an objective standpoint, since the BOP is less than the 30% threshold, I would classify this as localized."

Automatic CAL calculation function in axiUm may overestimate value of CAL in the presence of gingival inflammation and pseudo pockets

As described in Chapter 6.3, CAL is automatically entered in axiUm based on the value of PPD and the value of the gingival recession. This feature was entered in axiUm so that dental clinicians do not have to manually enter this information in axiUm. It was observed that in some cases, the CAL value automatically entered based on the PPD value and gingival recession may not be correct in the presence of gingival inflammation or pseudo pockets. A pseudo pocket is a pocket that results from gingival inflammation with edema that produces an apparent abnormal depth of the gingival sulcus without apical movement of the bottom of the sulcus; a false pocket. In the literature, there are two ways of measuring CAL; 1) direct method, and 2) indirect method. In the direct method, CAL is measured directly by the visualization of the probe over the reference point. In the indirect method, CAL value is calculated by adding PPD value and gingival recession values. However, from a mathematical perspective, because both PPD and gingival recession are subject to measurement errors, their combined use to determine the CAL could lead to the compounding of errors that could interfere with the final CAL value. As described earlier,

the PPD and gingival recession rates vary by intraoral conditions, such as inflammatory status, the presence of supragingival and subgingival calculus and tooth position in the arch (Barbosa et al, 2016, Corriaini, 2013, Bulthuis, 1998). In the following paragraph, one example from the manual review process. This patient did not have a radiographic bone loss; however, the CAL was recorded over 4 mm on two different teeth. Therefore, according to the case definitions, this patient was classified into one of the periodontitis cases.

Example comment from a manual reviewer:

> "Not enough information given. Will need to see radiographs to come up with a diangosis. The bone loss mentioned in the radiographic description may be due to physiologic bone loss. Furthermore, the deep probing depths may be due to pseudopocketing and gingival inflammation, rather than true CAL that can be directly attributable to periodontitis"

Two percent of same clinical note provided contraindicatory description of periodontal health.

During the manual review, it was also found that clinical notes recorded on the same date for the same patient had contraindicatory information. For example, information confirming that the patient is being healthy (coral pink gingiva, stippling present, etc) and having gingivitis (erythematous gingivitis, edematous, etc) was present in the same clinical note. There were three clinician notes (2%) out of 125 (100%) that had contraindicatory information. For these cases, reviewers requested actual radiographies of these patients. However, getting access of the radiographies is out of the scope of this study.

Example comments from a reviewer*:*

> "To be honest, I found this case to be confusing for two reasons. First, in the clinical description, the student provider writes that the mandibular anterior gingiva has knife-edged borders (which is an indication of healthy) with slightly bulbous gingival margins (which is an indication of disease). Second, in the radiographic section, the student provider indicates the presence of angular defects around #31. But in a subsequent sentence, he/she also writes that the bone levels is normal between 1-2 mm. These two sentences are in complete contradiction to each other. Also, probing depths and CAL are identical. So, it doesn't seem like there is any clinical attachment LOSS. Therefore, I found it very difficult to justify that this case's diagnosis is periodontitis. Also, I don't feel comfortable calling this a healthy gingiva (even though that's why I diagnosed the patient with this) because there is still clinical evidence of inflammation in the mandibular molar regions. Rather, I would feel much more comfortable if I could give the patient a diagnosis of localized gingivitis around #18 and 31 based on the clinical description of inflammation being present in those areas."

5.9: Prevalence of Periodontal Disease in IUSD Patient Population

The prevalence of PD was calculated using the clinician-recorded diagnoses. As shown in Table 31, it was observed that younger adults had significantly higher gingivitis than older individuals. Older individuals had significantly high periodontitis compared to younger individuals. Table 32 shows that female patients had significantly higher gingivitis than male patients and male patients had significantly high periodontitis than the female patients. By examining PD prevalence by race, Caucasians had higher gingivitis compared to African Americans. African Americans had higher periodontitis than Caucasians (see Table 33). It was also observed that patients who did not have dental insurance and self-paid for the treatments had significantly higher gingivitis and periodontitis compared to privately owned dental insurances (see Table 34).

Table 30: Prevalence of gingivitis and periodontitis based on clinician-recorded diagnoses by age groups at Indiana University School of Dentistry

	Healthy (%)	Gingivitis (%)	Periodontitis (%)	Total (%)	p value
18 to 29	39	2,123	203	2,365	<0.001
30 to 44	54	1,960	1,227	3,241	<0.001
45 to 64	143	2,019	2,967	5,129	<0.001
65 years or older	50	720	1,522	2,292	<0.001
Total	286	6,822	5,919	13,027	<0.001

Table 31: Prevalence of gingivitis and periodontitis by gender at Indiana University School of Dentistry

Gender	Healthy (%)	Gingivitis (%)	Periodontitis (%)	Total (%)	p value
Female	157 (54)	3,903 (57)	2,884 (48)	6,944 (53)	<0.001
Male	134 (46)	2,933 (43)	3,159 (52)	6,226 (47)	<0.001
Total	291 (2)	6,836 (52)	6,043 (46)	13,170 (100)	NA

Table 32: Prevalence of gingivitis and periodontitis by race at Indiana University School of Dentistry

Race	Healthy (%)	Gingivitis (%)	Periodontitis (%)	Total (%)	p value
Caucasians	155 (69)	3,713 (70)	2,950 (61)	6,818 (66)	<0.001
African Americans	55 (24)	1,041 (20)	1,349 (28)	2,445 (24)	<0.001
Other	16 (7)	535 (10)	538 (11)	1,089 (11)	<0.001
Total	266 (3)	5,289 (51)	4,837 (47)	10,352 (100)	NA

Table 33: Prevalence of gingivitis and periodontitis by insurance status at Indiana University School of Dentistry

Insurance	Healthy (%)	Gingivitis (%)	Periodontitis (%)	Total (%)	p value
Self-pay	153 (52)	3445 (50)	3176 (53)	6774 (51)	<0.001
Private	116 (40)	2919 (43)	2355 (39)	5390 (41)	<0.001
Government	24 (8)	501 (7)	528 (9)	1053 (8)	<0.001

5.10: Observation Time

When calculating patients' years of observation time using procedure codes related to periodontal treatments (COE, POE, PM, PRE), 0 to 15 years of observation time was found. Total of 15,217 (53%) patients out of 28,908 (100%) having no follow-up visits, 9,954 (34%) out of 28,908 (100%) patients having up to 5 years of observation time, 3,203 (11%) patients out of 28,908 (100%) having 5 years to 10 years of observation time, and 534 (2%) patients out of 28,908 (100%) having 10 years to 15 years of observation time (see Table 35) was observed.

Table 34: Table showing the number (%) of patients by the observation time between the first and last visits from June 1, 2005 to August 1, 2019 (COE, POE, PM, PRE)

Time in years (Observation time)	N	(%)
No follow-up	15,217	(53)
Up to 5 years	9,954	(34)
>5 and <=10 years	3,203	(11)
>10 and <=15 years	534	(2)
Total	28,908	(100)

Note: This table is generated using patients' comprehensive oral evaluation, periodic oral evaluation, periodontal maintenance, and periodontal reevaluation procedure code.

When calculating patients' years of observation time using their completed periodontal charting data, 0 to 15 years of observation time was found. Total 10,521 (37%) patients out of 28,908 (100%) having no follow-up visits, 9,651 (33%) out of 28,908 (100%) patients having up to 5 years of observation time, 2,322 (8%) patients out of 28,908 (100%) having 5 years to 10 years of observation time, and 386 (1%) patients out of 28,908 (100%) having 10 years to 15 years of observation time (see Table 36) was observed.

Table 35: Table showing the number (%) of patients by the observation time between the first and last visits from June 1, 2005 to August 1, 2019 (periodontal charting data)

Time in years (Observation time)	Frequency	(%)
No follow-up	10,521	(37)
Up to 5 years	9,651	(33)
>5 and <=10 years	2,322	(8)
>10 and <=15 years	386	(1)
>15 and <=20 years	0	(0)
Missing data	6,028	(21)
Total	28,908	(100)

The average number of clinician-recorded diagnoses available per patient who had at least two clinician-recorded diagnosis between June 1, 2005, to August 1, 2019 based on the documentation of two periodontal variables: periodontal charting findings and clinician documented PD diagnosis was also determined. For the periodontal charting findings, 2.78 average (median=2, standard deviation=2.9) documented charting findings for the patients who received COE, between January 1, 2009 to December 31, 2014 and if these patients received other treatments (such as COE, POE, PM, PRE) before 2009 or after 2014 (June 1, 2005, to August 1, 2019) (see Table 37) was observed. There were total of 63,552 periodontal charts documented for 22, 880 unique patients.

Examining documentation of clinician-recorded diagnoses in periodontal evaluation forms, 20,152 clinician-recorded diagnosis for 13,219 unique patients were found. As shown in Table 36, the average documented PD diagnosis was 1.52 (median=1, standard deviation=1) for 13,114 unique patients who received at least one COE between January 1, 2009 to December 31, 2014 and if these patients received other treatments before 2009 or after 2014 (June 1, 2005, to August 1, 2019). It was found that 7,657 (58%)

of patients had exclusively one clinician-recorded PD diagnosis, 3, 197 (24%) had exclusively two clinician-recorded PD diagnosis, 1,052 (8%) had exclusively three clinician-recorded diagnoses and 1,313 (10%) patients had 4 to 28 clinician-recorded PD diagnosis. There were total 5,562 patients who had more than one clinician-recorded diagnosis available to determine their disease change between their first and last visits (see Table 38).

Among the 5,562 patients who had more than two clinician-recorded diagnoses available, the average time gap between their first and second visit was 0.9 year (approximately 11 months (346 days)) (standard deviation of 584 days); first and third visit was 1.6 years (approximately 19 months (588 days)) (standard deviation of 709 days); and first and fourth visit was 3 years (approximately 35 months (1,072 days)) (standard deviation of 855 days).

Table 36: Descriptive statistics of patients' longitudinal periodontal charting information who received COE between January 1, 2009 to December 31, 2014 and received any other treatments between June 1, 2005, to August 1, 2019

Average visit	2.78
Median visit	2
Standard deviation	2.9
Minimum visit	1
Maximum visits	38
Total charts (multiple patients)	63,552
Unique patients	28,908 (100%)

Table 37: Descriptive statistics of patients' longitudinal clinician-documented periodontal disease diagnosis who received COE between January 1, 2009 to December 31, 2014 and received any other treatments between June 1, 2005, to August 1, 2019

Average visit	1.52
Median visit	1
Standard deviation	1
Minimum visit	1
Maximum visits	14
Total eval. forms	20,152
Unique patients	13,219 (46%)
Missing eval. forms	15,689 (54%)
Total unique patients	28,908 (100%)

5.11: Change in Patients' Periodontal Disease Status Over Time

Tables 41, 42, 43 demonstrate the number of patients whose PD status did not change over time, progressed from less severe condition to more severe condition or their disease status improved. Seventy two percent patients (3,919) out of 5,562 (100%) patients who had more than one clinician-recorded diagnoses between their first and last visit and did not have a disease status change between their first and last visits. One possible reason could be because of the periodontal treatments received by these patients which could be preventing the progression of the disease to more severity stages. Another possible reason could be because patient visits were clustered close to their initial COE date. However, due to the chronic nature of PD, its progression is slow and patients' PD disease status change may not occur so quickly.

There were 669 (13%) patients out of 5,562 (100%) patients (with more than one dental visit) whose disease status progressed between their first and last visit (see Table 35). The top three categories in disease progression included 1) progression from generalized mild periodontitis to localized moderate periodontitis (77 (12%) out of 669 (100%)), 2) progression from generalized moderate periodontitis to localized severe periodontitis (66 (10%) out of 669 (100%)), and 3) generalized mild periodontitis to generalized moderate periodontitis (56 (9%) out of 669 (100%)). It was observed that 589 (11%) patients out of 5,562 (100%) patients whose disease improved between their first and last visits (see Table 36). The top three categories in disease improvement included: 1) from generalized moderate periodontitis to generalized mild periodontitis (76 (13%) out of 537 (100%), 2) generalized mild periodontitis to generalized mild gingivitis (32 (5%) out

of 537 (100%), and 3) generalized mild periodontitis to localized mild periodontitis (30 (5%) out of 537 (100%)).

Table 40 shows the categories which were unknown (not able to determine disease change over time). Most of these categories either did not have information about a patient's disease type (gingivitis/periodontitis) or their disease severity (mild, mild to moderate, moderate, moderate to severe, severe). There were total 437 (7%) patients out of 5,486 (100%) patients in the unknown category.

Table 38: Number of patients whose disease status did not change from their first visit to their last visit between June 1, 2005 and August 1, 2019

From “disease stage” TO “disease stage”	Number of patients
generalized mild periodontitis to generalized mild periodontitis	38
generalized mild gingivitis to generalized mild gingivitis	729
generalized moderate periodontitis to generalized moderate periodontitis	725
generalized moderate gingivitis to generalized moderate gingivitis	439
generalized gingivitis to generalized gingivitis	274
localized mild periodontitis to localized mild periodontitis	215
generalized mild to moderate periodontitis to generalized mild to moderate periodontitis	140
generalized severe periodontitis to generalized severe periodontitis	90
localized moderate periodontitis to localized moderate periodontitis	84
localized mild gingivitis to localized mild gingivitis	80
generalized moderate to severe periodontitis to generalized moderate to severe periodontitis	73
generalized mild to moderate gingivitis to generalized mild to moderate gingivitis	48
generalized periodontitis to generalized periodontitis	42
localized gingivitis to localized gingivitis	39
localized severe periodontitis to localized severe periodontitis	34
localized periodontitis to localized periodontitis	30
generalized severe gingivitis to generalized severe gingivitis	29
generalized mild gingivitis to generalized mild periodontitis	27
localized mild periodontitis to generalized mild periodontitis	21
localized moderate gingivitis to localized moderate gingivitis	11
localized mild to moderate periodontitis to localized mild to moderate periodontitis	11
generalized moderate to severe gingivitis to generalized moderate to severe gingivitis	9
localized severe gingivitis to localized severe gingivitis	4
localized moderate to severe periodontitis to localized moderate to severe periodontitis	4
localized mild to moderate gingivitis to localized mild to moderate gingivitis	3

localized moderate to severe gingivitis to localized moderate to severe gingivitis	1
Total	3,919

Table 39: Number of patients whose disease status progressed from their first visit to the last visit between June 1, 2005 and August 1, 2019

From “disease stage” TO “disease stage”	Number of patients
generalized mild periodontitis to localized moderate periodontitis	77
generalized moderate periodontitis to localized severe periodontitis	66
generalized mild periodontitis to generalized moderate periodontitis	56
generalized gingivitis to localized mild periodontitis	35
generalized mild to moderate periodontitis to generalized moderate periodontitis	26
generalized mild periodontitis to localized severe periodontitis	26
generalized moderate gingivitis to localized mild periodontitis	25
localized mild periodontitis to generalized mild gingivitis	24
generalized moderate gingivitis to generalized mild periodontitis	23
generalized mild gingivitis to localized moderate periodontitis	18
generalized mild periodontitis to generalized mild to moderate periodontitis	17
generalized mild gingivitis to generalized moderate gingivitis	15
generalized moderate gingivitis to generalized moderate periodontitis	15
localized mild gingivitis to generalized mild gingivitis	14
generalized mild to moderate periodontitis to localized severe periodontitis	14
generalized gingivitis to localized moderate periodontitis	13
generalized mild to moderate periodontitis to localized moderate periodontitis	11
localized mild periodontitis to localized moderate periodontitis	11
generalized mild gingivitis to localized severe periodontitis	10
generalized moderate periodontitis to generalized severe periodontitis	10
generalized gingivitis to generalized mild periodontitis	9
generalized moderate gingivitis to localized severe periodontitis	7
generalized gingivitis to localized periodontitis	7
localized mild periodontitis to localized severe periodontitis	6
localized mild periodontitis to generalized moderate periodontitis	6
generalized mild gingivitis to generalized moderate periodontitis	6
generalized gingivitis to localized severe periodontitis	6
generalized mild gingivitis to localized moderate gingivitis	6
generalized severe gingivitis to localized mild periodontitis	5
generalized mild gingivitis to generalized mild to moderate gingivitis	5

generalized moderate to severe periodontitis to localized severe periodontitis	5
generalized mild periodontitis to localized moderate to severe periodontitis	4
localized mild gingivitis to localized mild periodontitis	4
generalized moderate gingivitis to generalized mild to moderate periodontitis	3
localized mild to moderate periodontitis to generalized mild periodontitis	3
generalized mild to moderate gingivitis to localized mild periodontitis	3
generalized moderate gingivitis to generalized mild to moderate gingivitis	3
generalized mild periodontitis to localized mild to moderate periodontitis	3
localized moderate periodontitis to generalized moderate periodontitis	3
localized mild gingivitis to generalized moderate gingivitis	3
localized mild to moderate periodontitis to generalized moderate periodontitis	2
generalized mild periodontitis to generalized severe periodontitis	2
localized moderate periodontitis to localized severe periodontitis	2
generalized moderate to severe gingivitis to generalized moderate periodontitis	2
localized mild periodontitis to localized moderate to severe periodontitis	2
localized mild to moderate gingivitis to localized mild periodontitis	2
generalized moderate gingivitis to localized severe gingivitis	2
generalized mild to moderate gingivitis to generalized mild periodontitis	2
generalized mild to moderate periodontitis to generalized severe periodontitis	2
generalized moderate periodontitis to generalized moderate to severe periodontitis	2
generalized moderate periodontitis to localized moderate to severe periodontitis	2
generalized mild to moderate gingivitis to localized moderate periodontitis	2
generalized moderate to severe periodontitis to generalized severe periodontitis	2
localized mild to moderate gingivitis to generalized mild to moderate gingivitis	2
generalized severe gingivitis to generalized mild periodontitis	2
generalized mild to moderate gingivitis to generalized moderate gingivitis	2
localized mild periodontitis to localized mild to moderate periodontitis	2
generalized mild gingivitis to generalized mild to moderate periodontitis	2

localized gingivitis to generalized gingivitis	2
generalized mild to moderate gingivitis to localized mild to moderate periodontitis	2
localized gingivitis to localized periodontitis	1
localized moderate gingivitis to generalized moderate gingivitis	1
localized mild gingivitis to localized moderate gingivitis	1
generalized severe gingivitis to localized severe periodontitis	1
generalized moderate to severe gingivitis to localized mild to moderate periodontitis	1
localized moderate gingivitis to generalized moderate periodontitis	1
localized moderate gingivitis to localized moderate periodontitis	1
generalized mild periodontitis to generalized moderate to severe periodontitis	1
generalized mild gingivitis to generalized severe gingivitis	1
localized mild gingivitis to generalized moderate to severe gingivitis	1
localized mild periodontitis to generalized severe periodontitis	1
generalized moderate gingivitis to generalized severe gingivitis	1
generalized mild gingivitis to localized mild to moderate gingivitis	1
localized severe periodontitis to generalized severe periodontitis	1
generalized mild to moderate gingivitis to generalized moderate periodontitis	1
generalized severe gingivitis to generalized mild to moderate periodontitis	1
generalized moderate to severe gingivitis to localized mild periodontitis	1
localized mild gingivitis to generalized moderate periodontitis	1
generalized mild to moderate periodontitis to generalized moderate to severe periodontitis	1
localized severe gingivitis to localized mild periodontitis	1
localized mild periodontitis to generalized mild to moderate periodontitis	1
localized moderate to severe gingivitis to generalized moderate periodontitis	1
localized moderate to severe gingivitis to generalized moderate to severe gingivitis	1
generalized severe gingivitis to localized moderate periodontitis	1
generalized moderate gingivitis to generalized moderate to severe periodontitis	1
generalized mild to moderate gingivitis to generalized moderate to severe gingivitis	1
Total	669

Table 40: Number of patients whose disease status improved from their first visit to the last visit between June 1, 2005 and August 1, 2019

From “disease stage” TO “disease stage”	Number of patients
generalized moderate periodontitis to generalized mild periodontitis	76
generalized mild gingivitis to localized mild periodontitis	50
generalized mild periodontitis to generalized mild gingivitis	32
generalized mild periodontitis to localized mild periodontitis	30
generalized mild gingivitis to localized mild gingivitis	20
generalized severe periodontitis to generalized moderate periodontitis	19
generalized moderate periodontitis to localized moderate periodontitis	19
generalized moderate gingivitis to localized moderate periodontitis	18
generalized moderate to severe periodontitis to generalized moderate periodontitis	18
generalized moderate gingivitis to generalized mild gingivitis	17
generalized mild periodontitis to generalized moderate gingivitis	15
generalized moderate periodontitis to generalized mild to moderate periodontitis	12
generalized moderate periodontitis to localized mild periodontitis	12
localized moderate periodontitis to generalized moderate gingivitis	11
generalized moderate periodontitis to generalized moderate gingivitis	10
generalized moderate gingivitis to localized mild gingivitis	10
generalized moderate periodontitis to generalized mild gingivitis	10
generalized severe periodontitis to localized severe periodontitis	9
localized moderate periodontitis to generalized mild periodontitis	8
localized No disease to localized No disease	7
localized mild periodontitis to localized mild gingivitis	7
generalized mild periodontitis to localized mild gingivitis	7
generalized mild periodontitis to localized moderate gingivitis	6
generalized moderate to severe periodontitis to generalized mild periodontitis	6
generalized gingivitis to localized mild gingivitis	5
localized moderate periodontitis to localized mild periodontitis	5
generalized moderate gingivitis to localized moderate gingivitis	4
generalized severe gingivitis to generalized mild gingivitis	4
generalized moderate to severe periodontitis to localized moderate to severe periodontitis	3
generalized severe periodontitis to generalized mild periodontitis	3

localized mild to moderate periodontitis to generalized mild to moderate gingivitis	3
generalized mild to moderate periodontitis to generalized mild to moderate gingivitis	3
generalized moderate periodontitis to generalized severe gingivitis	3
localized mild to moderate periodontitis to localized mild periodontitis	3
localized severe periodontitis to generalized moderate periodontitis	3
localized severe periodontitis to localized mild periodontitis	3
generalized mild to moderate gingivitis to generalized mild gingivitis	3
generalized severe periodontitis to generalized mild to moderate periodontitis	2
generalized mild periodontitis to generalized severe gingivitis	2
generalized mild to moderate periodontitis to localized mild periodontitis	2
generalized mild to moderate periodontitis to generalized moderate gingivitis	2
generalized moderate to severe periodontitis to generalized mild to moderate periodontitis	2
generalized mild to moderate periodontitis to generalized mild gingivitis	2
localized moderate to severe periodontitis to generalized moderate periodontitis	2
localized moderate to severe periodontitis to generalized mild periodontitis	2
generalized gingivitis to localized gingivitis	2
generalized mild to moderate periodontitis to generalized severe gingivitis	2
localized severe periodontitis to generalized mild periodontitis	2
localized moderate periodontitis to generalized mild gingivitis	2
localized moderate to severe periodontitis to generalized moderate gingivitis	2
generalized severe periodontitis to generalized moderate to severe periodontitis	2
localized moderate to severe periodontitis to generalized moderate to severe gingivitis	1
generalized moderate periodontitis to localized mild gingivitis	1
generalized moderate to severe periodontitis to localized moderate periodontitis	1
localized moderate to severe gingivitis to localized moderate gingivitis	1
generalized moderate periodontitis to localized moderate gingivitis	1

generalized moderate periodontitis to generalized mild to moderate gingivitis	1
localized mild periodontitis to localized severe gingivitis	1
localized moderate to severe periodontitis to generalized mild No disease	1
generalized severe periodontitis to generalized severe gingivitis	1
localized severe periodontitis to generalized mild to moderate gingivitis	1
localized moderate periodontitis to localized severe gingivitis	1
generalized periodontitis to generalized gingivitis	1
generalized mild periodontitis to localized severe gingivitis	1
localized mild periodontitis to localized moderate gingivitis	1
generalized mild to moderate periodontitis to localized mild to moderate periodontitis	1
localized moderate periodontitis to generalized severe gingivitis	1
generalized moderate to severe periodontitis to localized mild periodontitis	1
generalized severe periodontitis to localized mild to moderate periodontitis	1
localized moderate periodontitis to generalized mild to moderate gingivitis	1
generalized severe periodontitis to localized mild periodontitis	1
localized severe periodontitis to localized mild to moderate periodontitis	1
localized moderate to severe periodontitis to generalized severe gingivitis	1
generalized severe periodontitis to generalized moderate gingivitis	1
generalized moderate to severe gingivitis to generalized mild gingivitis	1
localized moderate periodontitis to localized moderate gingivitis	1
generalized severe gingivitis to localized mild gingivitis	1
generalized severe gingivitis to generalized mild to moderate gingivitis	1
localized periodontitis to generalized gingivitis	1
localized moderate gingivitis to generalized mild gingivitis	1
generalized gingivitis to localized mild to moderate gingivitis	1
localized gingivitis to localized severe gingivitis	1
localized moderate to severe periodontitis to generalized mild gingivitis	1
localized severe periodontitis to generalized severe gingivitis	1
localized mild periodontitis to generalized severe gingivitis	1
generalized moderate to severe periodontitis to generalized moderate gingivitis	1
Total	537

Table 41: Unknown periodontal disease change categories for which either disease type or severity information was not available from clinician-recorded diagnoses

FROM disease type TO disease type	Number of patients
generalized mild No disease to generalized mild No disease	46
generalized moderate No disease to generalized moderate No disease	40
generalized gingivitis to generalized mild gingivitis	30
generalized mild to moderate periodontitis to generalized mild periodontitis	22
generalized moderate periodontitis to generalized moderate No disease	15
generalized No disease to generalized No disease	15
generalized mild gingivitis to localized mild No disease	9
generalized periodontitis to generalized moderate periodontitis	8
localized mild periodontitis to generalized moderate gingivitis	8
generalized mild periodontitis to localized moderate No disease	8
generalized gingivitis to generalized mild to moderate gingivitis	7
generalized moderate periodontitis to localized severe No disease	7
generalized mild to moderate periodontitis to generalized mild No disease	6
generalized mild to moderate No disease to generalized mild to moderate No disease	6
generalized gingivitis to generalized moderate gingivitis	6
generalized mild periodontitis to localized severe No disease	5
generalized mild periodontitis to generalized moderate No disease	5
generalized mild No disease to generalized mild periodontitis	5
localized mild No disease to localized mild No disease	5
localized moderate No disease to localized moderate No disease	5
generalized gingivitis to localized No disease	4
generalized mild periodontitis to generalized mild No disease	4
generalized moderate gingivitis to localized moderate No disease	4
generalized moderate periodontitis to localized moderate No disease	4
generalized gingivitis to generalized severe gingivitis	3
generalized gingivitis to generalized moderate periodontitis	3
generalized mild gingivitis to generalized moderate No disease	3
generalized moderate gingivitis to generalized moderate No disease	3
generalized mild periodontitis to localized mild No disease	3
generalized moderate No disease to generalized moderate periodontitis	3
generalized periodontitis to localized moderate periodontitis	3
localized gingivitis to generalized mild periodontitis	3

generalized moderate No disease to localized severe periodontitis	3
generalized mild No disease to generalized moderate periodontitis	3
localized gingivitis to generalized mild gingivitis	3
generalized moderate periodontitis to generalized mild No disease	3
generalized periodontitis to generalized mild periodontitis	3
generalized mild gingivitis to localized moderate No disease	3
generalized gingivitis to localized mild No disease	3
generalized periodontitis to localized severe periodontitis	3
generalized periodontitis to localized mild periodontitis	3
generalized mild No disease to generalized mild gingivitis	3
localized mild periodontitis to localized mild No disease	2
localized mild gingivitis to generalized mild No disease	2
localized gingivitis to generalized moderate gingivitis	2
localized severe periodontitis to localized severe No disease	2
generalized mild to moderate periodontitis to generalized mild to moderate No disease	2
generalized mild gingivitis to generalized mild No disease	2
generalized moderate periodontitis to localized mild No disease	2
generalized moderate No disease to generalized mild periodontitis	2
localized moderate periodontitis to generalized mild No disease	2
generalized periodontitis to generalized severe periodontitis	2
localized mild periodontitis to generalized mild No disease	2
generalized mild to moderate periodontitis to generalized moderate No disease	2
localized mild No disease to generalized mild gingivitis	2
generalized moderate to severe periodontitis to generalized moderate No disease	2
localized mild gingivitis to localized mild No disease	2
localized periodontitis to generalized moderate periodontitis	2
generalized mild to moderate gingivitis to generalized mild to moderate No disease	2
generalized No disease to generalized mild to moderate periodontitis	2
generalized moderate gingivitis to localized mild No disease	2
generalized No disease to generalized mild gingivitis	2
localized periodontitis to localized mild periodontitis	2
generalized mild No disease to localized moderate periodontitis	2
generalized gingivitis to localized moderate No disease	1
generalized mild to moderate No disease to localized mild to moderate periodontitis	1

generalized moderate to severe periodontitis to generalized mild gingivitis	1
generalized mild No disease to localized moderate gingivitis	1
localized No disease to generalized mild periodontitis	1
generalized moderate No disease to generalized severe periodontitis	1
localized moderate gingivitis to localized severe No disease	1
localized moderate No disease to localized moderate periodontitis	1
generalized periodontitis to localized No disease	1
generalized moderate No disease to localized severe No disease	1
localized mild to moderate periodontitis to localized severe No disease	1
generalized severe periodontitis to localized moderate to severe No disease	1
localized moderate periodontitis to localized moderate No disease	1
generalized mild No disease to localized severe periodontitis	1
generalized periodontitis to localized moderate No disease	1
localized periodontitis to localized severe periodontitis	1
generalized No disease to generalized mild periodontitis	1
generalized moderate periodontitis to generalized severe No disease	1
generalized mild No disease to generalized moderate gingivitis	1
generalized mild gingivitis to localized severe No disease	1
localized No disease to generalized gingivitis	1
localized moderate No disease to localized severe periodontitis	1
generalized moderate No disease to generalized mild gingivitis	1
generalized No disease to localized mild to moderate periodontitis	1
generalized periodontitis to generalized mild gingivitis	1
generalized periodontitis to generalized severe No disease	1
generalized severe No disease to generalized moderate No disease	1
generalized gingivitis to generalized No disease	1
generalized moderate periodontitis to generalized mild to moderate No disease	1
generalized moderate to severe periodontitis to generalized mild No disease	1
generalized moderate No disease to localized moderate gingivitis	1
localized mild periodontitis to localized severe No disease	1
generalized periodontitis to localized moderate to severe periodontitis	1
localized mild periodontitis to generalized mild to moderate No disease	1
generalized moderate to severe periodontitis to localized moderate to severe No disease	1
generalized mild No disease to localized moderate No disease	1
generalized periodontitis to localized severe No disease	1
generalized mild to moderate gingivitis to generalized mild No disease	1

generalized moderate to severe periodontitis to localized moderate No disease	1
generalized mild to moderate gingivitis to localized mild No disease	1
generalized severe No disease to generalized mild periodontitis	1
localized No disease to generalized mild No disease	1
localized gingivitis to localized mild periodontitis	1
localized gingivitis to localized No disease	1
localized gingivitis to localized moderate periodontitis	1
generalized gingivitis to generalized mild to moderate periodontitis	1
generalized moderate to severe No disease to generalized moderate to severe No disease	1
generalized mild gingivitis to localized mild to moderate No disease	1
generalized severe periodontitis to generalized moderate No disease	1
generalized periodontitis to generalized mild to moderate periodontitis	1
localized periodontitis to generalized mild gingivitis	1
localized periodontitis to generalized moderate gingivitis	1
localized gingivitis to localized severe periodontitis	1
generalized severe No disease to generalized severe No disease	1
localized periodontitis to generalized moderate No disease	1
generalized mild periodontitis to localized moderate to severe No disease	1
generalized gingivitis to localized moderate gingivitis	1
generalized mild No disease to localized mild periodontitis	1
generalized moderate gingivitis to generalized mild No disease	1
generalized No disease to localized severe periodontitis	1
localized moderate to severe periodontitis to localized severe No disease	1
generalized moderate to severe No disease to generalized severe No disease	1
localized mild periodontitis to localized moderate No disease	1
generalized moderate to severe No disease to generalized moderate periodontitis	1
localized No disease to localized periodontitis	1
Total	437

6: Discussion

The objective of the study was to determine the quality of the clinical findings that characterize PD diagnoses in the EDR and to test the feasibility of tracking PD status change over time. The study's goal was to determine how complete the periodontal data is documented in the EDR and if diagnoses determined from findings were similar to clinician-recorded diagnoses to automatize the process of generating PD diagnoses. The use of longitudinal EDR data to study disease progression was also determined in this study. Last, the prevalence of gingivitis and periodontitis using the EDR data was calculated. This chapter will begin with a discussion of the approach used to achieve the objective, followed by motivation of this research, findings from both of the aims, and then the comparison of the study results with previous studies. Last, this section describes the limitation of the study and the final conclusion of this research.

6.1: Motivation for Conducting This Study

It is well understood that there is an increased adoption of electronic dental records (EDR) from the past decade to document patient care information electronically [110, 115]. As a result, there is a growing interest in utilizing patient care information stored in the EDR for research purposes [25] because EDR data offers an opportunity to characterize current patient populations. Therefore, many researchers have utilized EDR data to assess various dental treatment outcomes such as non-surgical root canal treatment [21, 38, 59, 97, 121, 122], longevity of crown [22], and posterior composites restorations [74]. Despite the promising potential of EDR data for research, they come with their own challenges such as questionable quality of data, and missing data. Therefore, before utilizing EDR data for research, it is import to first determine the quality of the EDR data and whether it could be used for research purposes. Only one study exists by Thyvalikakath et al [120] that developed data quality metrics to assess the quality of the EDR data. No study exists to determine the completeness of data to study PD using EDR data, the process of generating a cohort of PD patients and the challenges involved. In summary, it is critical to first evaluate the quality of the EDR data before its use because flawed data could result in flawed outcomes/results. In this study, a cohort of PD patients was generated and investigated up to what extent the EDR data can be used for clinical research.

6.2: Main Highlights of the Study

This study provided the groundwork to evaluate the quality of the periodontal findings and diagnosis of PD information stored in the EDR before its use for clinical research and a process of utilizing periodontal findings to generate PD diagnoses. To the best of the knowledge, this was the first study that utilized patients' BOP information from the periodontal charting data and classified their gingival health based on the BOP score. This was also the first study which utilized all six interproximal sites to determine patients' periodontitis diagnosis from periodontal findings that were limitations of the NHANES periodontitis prevalence study [33].

Excellent data quality was observed for patient demographic variables such as date of birth, gender, and insurance information which were recorded for all patients (28,908 (100%)), periodontal charting data was available for 80% (22, 880) of patients, moderate data quality for race information (completeness of 72% (20,880)), and clinician-recorded diagnoses (completeness of 46% (13,219)). Due to the cut-off criteria of the BOP score, some of the gingivitis cases were missed compared to the clinician-recorded diagnoses. Upon manual review, 30% (37 records out of 125) of patients' gingival and intra oral examination findings demonstrated the presence of gingivitis, however, the cut off criteria of the BOP score wasn't 10% or more to diagnose these cases in any gingivitis categories. Looking into periodontitis cases, there were significantly higher cases when generated from periodontal findings compared to clinician-recorded diagnoses.

When utilizing longitudinal EDR data to track a patient's disease change, it was found that out of 13,219 patients (who had at least one clinician-recorded diagnosis available), only 5,486 (42%) patients had information available for more than one visit

allowing this study to track the disease progression. We found majority of patient information clustered in the beginning of the study period, as a result, 70% of the patient population was found to be in no disease status change category while examining their longitudinal EDR data. This study also found that 3,949 (72%) patients out of 5,486 (100%) did not experience a disease status change between their first and last visits. One possible reason could be because of the periodontal treatments received by these patients which is preventing the progression of the disease to more severity stages. Advances have been made in periodontal treatments over the last three decades and the provided treatments could be helping patients to arrest the disease by preventing its progression. It could be also because patients' visits were clustered at the beginning of the observation study period and due to the slowly progressing nature of PD the disease change may not occur quickly.

6.3: Case Definitions of Periodontal Disease

The 2018 study [33] that examined the prevalence of periodontitis in the US population has used the case definitions developed by the CDC and AAP to determine patients' periodontitis prevalence. These case definitions are described in Chapter 2.3. The authors also measured the extent of the PD by measuring CAL and PPD on six sites per tooth. However, while calculating the prevalence of the PD, they used measurements from four interproximal sites (mesiobuccal, distobuccal, mesiolingual, distolingual) with an assumption that those sites are most affected by the disease and excluded midbuccal and midlingual sites. Measurements from the mid-buccal and the mid-lingual sites that potentially could indicate furcation involvement were not included in the study. In addition, they also excluded involvement of people for medical reasons and people who are institutionalized such as nursing home residents which may have introduced selection bias. Because of time constraints the examiners did not assess bleeding on probing sites which could provide information to estimate gingivitis prevalence which is a precursor of periodontitis [33]. The study authors acknowledged that they may have had underestimated the disease prevalence.

In contrast to this nationwide prevalence study, this study used all six sites per tooth which helped in estimating the health of the entire tooth. Moreover, no patients were excluded based on their medical conditions or institutionalization which represented a real world patient population. Since patients' BOP information is recorded in the periodontal charting findings, patients' gingivitis status was measured based on the case definition demonstrated in the study [126].

6.4: Reasons for Low Agreement

While comparing agreement between diagnoses generated from the periodontal findings and clinician-recorded diagnoses, there was a 47% agreement. Below, the possible reasons for the moderate to a low agreement are described.

Gingivitis cases

Significantly fewer gingivitis cases were found when comparing patients' gingivitis diagnosis determined using the BOP score with the clinician-recorded diagnosis. Based on the manual review, two possible reasons are suspected. First, the cut-off criteria of the BOP score used to define gingivitis may not be representative of patients' gingival health. As described in the results section, the reviewers found that out of 125 patients, 30% (39 patients) were diagnosed in the gingivitis category by the clinicians. However, the BOP score did not meet the 10% or more criteria to classify these patients into gingivitis cases. During the manual review process, the study discovered that these patients had gingival inflammation, edema and other signs of gingivitis recorded in their periodontal evaluation form. Reviewers and clinicians diagnosed these patients to have gingivitis, but according to the BOP score, these patients were classified as a healthy case which resulted in disagreements.

Periodontitis cases

There were a significantly higher number of patients with periodontitis cases when diagnosis was determined from periodontal findings rather than clinician-recorded diagnosis. It is possible that the disagreements are due to the parameters suggested by the case definition of periodontitis [32, 86]. The *periodontitis diagnose.py* algorithm used the case definition which is formulated based on only CAL and PPD parameters and does not

consider other important parameters such as radiographic bone loss during the diagnosis. In contrast, clinicians have diagnosed patients' periodontitis status using radiographic and intraoral findings. However, it is well understood that a relatively small change in CAL and PPD values can result in large changes in the periodontitis diagnosis and PD prevalence. CAL is an accurate measure to be used to diagnose a patient's periodontitis diagnosis, however, it is difficult to measure CAL because it could vary based on gingival inflammation, calculus, edema, and pseudo-pockets. It is also observed that the CAL value is rarely used in daily clinical practice to diagnose periodontitis and mostly used in epidemiological and clinical trial studies [12].

On the other hand, it is also possible that the prevalence determined in the Eke et al study is underestimated because the authors included only four interproximal sites and excluded third molars (P. I. Eke et al., 2018). In this study, all probing sites are included which may better represent the actual prevalence of periodontitis.

Next, during the study period from January 1, 2009, to December 31, 2014, clinicians at IUSD have used the criteria which were recommended by AAP [6, 10]. In the classification, authors have proposed to use the same CAL and PPD criteria as described in the case-definitions, in addition to the radiographic bone loss (see Chapter 2.2). Upon manual review, it was observed that dental clinicians reported higher values of CAL in the periodontal charting findings. Though, when the same patient's clinical notes were examined, clinicians mentioned no radiographic bone loss. As a result, they classified these patients as healthy cases. The radiographic bone loss information was not included while determining patients' periodontitis diagnoses and therefore, resulted in low agreement.

Another possible reason for the low agreement could be the mode of calculating CAL values in axiUm. As described earlier in Chapter 4.4, CAL value is calculated automatically based on the PPD and gingival recession values (PPD + gingival recession = CAL called the indirect method). However, studies have shown that CAL measured using the indirect method could lead to the measurement errors compared to CAL measured directly from subtracting the distance between the cementoenamel junction and the free gingival margin [12] (Barbosa, Angst, Finger Stadler, Oppermann, & Gomes, 2016). Therefore, in some cases due to inaccurate values of CAL, the periodontitis diagnoses may have been overestimated.

6.5: Periodontal Disease Change

As described in Chapter 6.2, clinician-recorded diagnoses documented from patients' all visits between June 2005 to August 2019 (who received at least one COE between January 1, 2009, and December 31, 2014) were retrieved. While determining these patients' PD progression, the study found two average visits per patient between June 2005 and August 2019. When patients' disease status change over time was assessed, most of the patients (70%) were found to fall under the "no disease change" group. There could be two possible reasons for this. Advances have been made in periodontal treatments over the last three decades and the provided treatments could be helping patients to arrest the disease by preventing its progression. Moreover, when examining these patients' disease severity, nearly all of them (97%) were in either mild or moderate disease stage which is manageable to control. However, once the disease enters in the severe case, it becomes difficult to control the progression. Studies have shown that if regular treatment is provided to a periodontitis patient and good oral hygiene is maintained, then the disease progression can be stopped, especially when periodontitis is in the early disease stage.

While determining these patients' PD progression, at least two clinician-recorded diagnoses were utilized per patient within the 14 years of observation period (June 1, 2005 and August 1, 2019). Even though the observation period spanned 14 years, patients' visits were clustered at the beginning of that period. Since, chronic PD is a slow progressing disease and the disease progression varies based on patient characteristics, this could be another possible explanation behind 70% of patients being in "no disease change" group. This does not mean that the patient did not visit back to IUSD for treatment, nevertheless, their clinician-recorded diagnoses were not available for all of their visits. Because for

many visits, patients' charting data was recorded, however, their clinician-recorded diagnoses were missing.

6.6: Comparing the Study Results with Other Studies

Data quality

Recently, Thyvalikakath et al evaluated completeness and correctness of the data required to perform survival analysis for two dental treatments 1) posterior composite restorations and 2) root canal treatments on permanent teeth. The authors utilized 99 dental practices EDR data across the US and found nearly 100% completeness of patients' date of birth, gender, insurance, and procedure variables. Similarly, 100% completeness of patients' date of birth, gender, insurance, and procedure code variables was observed in this study. In the study by Thyvalikakath et al, authors were not able to obtain patients' race information as it was not available in the EDR. In this study, race/ethnicity information was obtained for 72% of patients. Alwhaibi et al measured the completeness of medication-related information from EHR and found that for about 100% of patients, age and gender were reported, similar to this study's results.[5]. Their study did not examine the completeness of patients' ethnicity/race information. Kopcke et al evaluated completeness in the EHR for the purpose of patient recruitment into clinical trials [52]. Their study's results demonstrated that age and gender variables were recorded for 89% of their patients. Hegde et al developed a non-invasive diabetes risk prediction model for application in the dental clinical environment [44]. As a part of their study, they reported the percentage of missing information in the data and they found that patients' age and gender were reported for all the patients. Patients' race information was missing for 3% of their patient population. In contrast, race information was available for only 72% of patients in this study's patient population. This study found that patients' periodontal charting information was recorded for 80% of patients, while the clinician-recorded periodontal diagnosis is only available for

47% of patients. There is only one study [44] that evaluated the completeness of periodontal charting data and found that 80% of their patients' BOP were missing, and 45% of their periodontitis diagnosis were missing. In contrast, at IUSD, documenting patients' periodontal charting information is mandatory while conducting COE. Therefore, excellent quality of documenting periodontal charting data (80% completeness) was observed. Other studies that attempted to research PD using EDR data did not use patients' charting data to diagnose patients' periodontitis status. These studies have assessed x-rays manually and determined bone loss information from the radiographs. [19, 20, 98].

Quality of longitudinal EDR data for research

Thyvalikakath et al, 2019 evaluated the quality of longitudinal EDR data to perform survival analysis of two dental procedures 1) posterior composite restorations and 2) root canal treatments. The authors examined the availability of longitudinal EDR data and follow up visits. They found that 42% of patient records had at least five years of observation time, 22% of patients had 5 to 10 years of observation time, and 14% of patients had up to 15 years of observation time. Only 15% of patients did not have a follow-up visit after the initial date of performing the treatment. In contrast, in this study, only nearly 19% (5,562 of 28,908) of patients were found to have two PD diagnoses information available for more than one visit between June 1, 2005 and August 1, 2019. This could be because Thyvalikakath et al studied the treatment outcome of a procedure (using dental procedure codes) in which disease diagnosis information was not required. In this study, the progression of a disease was examined as opposed to treatment outcome, which requires the availability of a disease diagnosis during each patient visit. In dentistry, unlike medicine, dental clinicians do not require patients' diagnostic information to get

reimbursed from the insurance agencies. Dentists get reimbursed by submitting the procedure codes. As a result, they are excellent in recording each procedure performed accurately; however, since the diagnosis is not required, they may not document this information in the EDR.

Periodontal disease prevalence

There are a few studies published that examined the prevalence of periodontitis [1, 33]. Eke et al estimated the prevalence of periodontitis in the US population using the NHANES data. Eke et al estimated the prevalence of periodontitis in the US population using the NHANES data [33]. Authors found that 42% of adults aged 30 years and older had periodontitis in years 2009 to 2014. Their study sample size was 10,683 patients representing the entire US population. Acharya et al, 2014, studied PD prevalence in the Wisconsin population and also found that almost half of their population had periodontitis [1], consistent with study results [32]. In this study, the same criteria that was used in Eke et al, 2018, and Acharya et al, 2014 studies were used to estimate PD prevalence. However, in this study it was found that 55% of the patients had moderate and 27% patients had severe periodontitis, with less than 1% of patients having mild periodontitis which makes it a total of 83% of patients having periodontitis. This could be because this study dataset was generated from an academic setting where patients visit when they have severe problems that need immediate attention.

In the recently published epidemiological study [33] the authors found that periodontitis prevalence was mostly observed in the elderly population, male, Hispanic, and African American demographics. This study also found similar results that the older population, African American race, and male gender had a significantly higher prevalence.

It was also observed that females had a significantly higher rate of gingivitis than males and this could be because female hormones may increase plasma levels which leads to gingival inflammation [137].

6.7: Limitations

Like any other study, this study encountered some limitations.

First, patients' soft tissue and intraoral examination findings such as gingival color, gingival contour, stippling, and inflammation were not included in this study which are essential components in diagnosing patients' PD diagnosis.

Second, patients' radiographic findings were not included while diagnosing their periodontitis status. It was observed that only relying on patients' CAL and PPD information may not be sufficient and radiographic findings are required while diagnosing patients' accurate periodontitis status.

Third, patients' gingival recession and pseudo pocket information were not considered while determining their periodontitis status. Studies have indicated that [12, 24] CAL value can be overestimated in the presence of gingival inflammation and pseudo pockets.

Fourth, only completeness and concordance data quality measures were examined. However, other data quality measures such as accuracy, plausibility, and reliability due to the lack of a gold standard dataset were not examined. The ideal approach to examine these measures would be to compare the reported information in the EDR with findings recorded directly from patients.

Fifth, the average follow-up visits were calculated from June 1, 2005, to August 1, 2019 time period. However, average patient visits by each year were not calculated which is important to determine how the patient visits are parsed across the time.

Last, the EDR dataset was utilized from only one institute, as a result, findings generated from one institute's dataset may not be generalizable.

6.8: Future Work

The results of this project provide a metrics to evaluate EDR data quality before its use for research and quality improvement purposes. It also provides a framework of generating a cohort of patients with PD from periodontal charting findings.

Future research should include patients' soft tissues, hard tissues, and intraoral findings to generate patients' gingivitis diagnoses because just relying on the BOP score isn't enough for diagnosing gingivitis status. Patients' soft tissues, hard tissues, and intraoral findings are recorded in a free-text format within the periodontal evaluation forms. Therefore, text-mining and natural language processing algorithms to extract this information to automate gingivitis diagnosis should be developed.

Utilizing radiographic findings information to determine patients' periodontitis status should also be considered. Similar to the soft tissue findings, patients' radiographic bone loss information is also recorded in the free-text format. Therefore, text-mining and natural language processing algorithms to extract this information automatically from the radiographic bone loss section of the periodontal evaluation form should be developed. Further research is required to examine the reliability of clinician-recorded radiographic findings in the periodontal evaluation forms by manually interpreting their radiographs, and also to explore the feasibility of extracting radiographic findings directly from patients' radiographs through image processing methods.

Further research should focus on evaluating longitudinal EDR data quality by calculating the information score described in [116], to examine the irregularity of temporal information present in the longitudinal EDR data. This will help researchers determine how longitudinal EDR data has been sparse as well as the variability of the time gaps between

observations. The information score is scaled from 0 to 1 for each observation, where 1 represents equally spaced and 0 represents sporadically placed. The lower values (less than 0.5) indicate the clusters near the beginning or at the end of the observation period.

7: Conclusions

There are numerous significant outcomes from this research study. First and foremost, this study demonstrated the significance of utilizing EDR data for PD research because EDR provided excellent data quality regarding patient demographics, insurance information, dental treatments, and periodontal charting. This is because, at the IUSD, the department of periodontology and allied dental health is the only dental school out of 66 schools in the nation that have been practicing calibration among dental faculties and students. Calibration practice at IUSD institute significantly improved the consensus and consistency of PD diagnosis and treatment planning. This study discovered moderate data quality of clinician-recorded PD diagnoses; however, the feasibility of automatically generating diagnosis from periodontal findings was tested, and achieved excellent performance. The results of this dissertation advise other dental schools to do regular calibrations and continuing education courses for good documentation for their faculty and students. This study discovered one limitation of the BOP score, the cut-off point that missed diagnosing 30% of the gingivitis patients. Therefore, to accurately assess patients' gingivitis diagnoses, either intraoral and soft tissue findings should be included along with the BOP score or the BOP cut off criteria (BOP score 10% or more) should be redefined. This was the first study that utilized patients' BOP information from the EDR data to generate gingivitis diagnosis. This study results also demonstrated the feasibility of estimating patients' periodontitis prevalence from the EDR data that has significant advantages over current approaches used in the epidemiological study assessing the prevalence of periodontitis in the US adult population. Compared to the epidemiological study, this study included all six sites per tooth, and all patient cases regardless of their

medical conditions. The results of this dissertation conclude that the longitudinal EDR data could be utilized to determine the short-term outcome of PD treatments using clinician-recorded diagnoses, and long-term outcome using the PD diagnoses generated from findings. This study demonstrated the feasibility of utilizing longitudinal EDR data to automatically detect patients' disease change over time. This information serves to increase awareness of the EDR use to determine PD treatment outcome, study clinical course of PD, and develop a prediction model that can represent the current patient population.

8: Proposed Publications

1. Effectiveness of 2018 gingivitis classification system in a clinical setting.
2. Agreement between periodontitis diagnoses generated from 2012 periodontitis case definition and clinicians' diagnoses.
3. Utilizing longitudinal electronic dental record data to track patients' periodontal disease change over time.
4. Estimating prevalence of gingivitis and periodontitis in Indiana population.

9: Appendix

This project was approved by the Indiana University Institutional Review Board (IRB). This study IRB protocol number is 1909819686. IRB approved this study as Exempt research.

Description: First, *"Charting Subset Extractor.py"* imported the *"pandas"* library which has functions for data manipulation and analysis. In this algorithm, the *"read_tsv"* function from this library to import and read the master dataset "Perio_Charting_Data.tsv" was used. *Charting Subset Extractor.py* used the *regular expression re ()* and the *find ()* functions (see Table 6) to search keywords for "ATTACH", and "POCKET" in the master dataset. Next, based on the presence of these keywords, this algorithm extracted patients' CAL, and PPD information from all charting information present in the master dataset. Only patients' CAL and PPD information was extracted because the case definition of periodontitis to determine patients' periodontitis status and these case definitions only required to use patients' CAL and PPD information. Last, this algorithm created an output file "Perio_Charting_Attach_Pocket_Subset. TSV" which contained patients' CAL and PPD information. Similarly, another subset file using the same algorithm that contains patients' only CAL and BOP information was created.

Table 41: Charting Data subset extractor algorithm to only extract patients' clinical attachment loss, periodontal pocket depth, and bleeding on probing information from master dataset

```
import pandas as pd
df = pd.read_tsv('PerioCHARTING_ALL.tsv', sep='\t')
output_file = open('PerioCHARTING_ATTACH_POCKET_Subset.tsv','w')
output_file.write('Id\tChartDate\tPerCond\tSiteName\tSection\tValue
1\tValue2\tValue3\n')
for i in range(len(df)):
    percond = df['PerCond'][i]
    if percond == 'ATTACH' or percond == 'POCKET':
        output_file.write(str(df['Id'][i]) + '\t' +
str(df['ChartDate'][i]) + '\t' + str(df['PerCond'][i]) + '\t' +
str(df['SiteName'][i])+ '\t' + str(df['Section'][i])+ '\t' +
str(df['Value1'][i])+ '\t' + str(df['Value2'][i])+ '\t' +
str(df['Value3'][i]) + '\n')
output_file.close()
output_file = open('PerioCHARTING_ATTACH_BLEED_Subset.tsv','w')
output_file.write('irb_id\tChartDate\tPerCond\tSiteName\tSection\tV
alue1\tValue2\tValue3\n')
for i in range(len(df)):
    percond = df['PerCond'][i]
    if percond == 'ATTACH' or percond == 'BLEED ':
        output_file.write(str(df['Id'][i]) + '\t' +
str(df['ChartDate'][i]) + '\t' + str(df['PerCond'][i]) + '\t' +
str(df['SiteName'][i])+ '\t' + str(df['Section'][i])+ '\t' +
str(df['Value1'][i])+ '\t' + str(df['Value2'][i])+ '\t' +
str(df['Value3'][i]) + '\n')
output_file.close()
```

Table 42: Charting Data subset abstractor algorithm to only extract patients' clinical attachment loss, periodontal pocket depth, and bleeding on probing information

```
import pandas as pd
df = pd.read_csv('PerioCHARTING_ALL.tsv', sep='\t')
output_file = open('PerioCHARTING_ATTACH_POCKET_Subset.tsv','w')
output_file.write('Id\tChartDate\tPerCond\tSiteName\tSection\tValue
1\tValue2\tValue3\n')
for i in range(len(df)):
    percond = df['PerCond'][i]
    if percond == 'ATTACH' or percond == 'POCKET':
        output_file.write(str(df['Id'][i]) + '\t' +
str(df['ChartDate'][i]) + '\t' + str(df['PerCond'][i]) + '\t' +
str(df['SiteName'][i])+ '\t' + str(df['Section'][i])+ '\t' +
str(df['Value1'][i])+ '\t' + str(df['Value2'][i])+ '\t' +
str(df['Value3'][i]) + '\n')
output_file.close()
output_file = open('PerioCHARTING_ATTACH_BLEED_Subset.tsv','w')
output_file.write('irb_id\tChartDate\tPerCond\tSiteName\tSection\tV
alue1\tValue2\tValue3\n')
for i in range(len(df)):
    percond = df['PerCond'][i]
    if percond == 'ATTACH' or percond == 'BLEED ':
        output_file.write(str(df['Id'][i]) + '\t' +
str(df['ChartDate'][i]) + '\t' + str(df['PerCond'][i]) + '\t' +
str(df['SiteName'][i])+ '\t' + str(df['Section'][i])+ '\t' +
str(df['Value1'][i])+ '\t' + str(df['Value2'][i])+ '\t' +
str(df['Value3'][i]) + '\n')
output_file.close()
```

Table 43: Clinician-recorded periodontal disease diagnosis extraction from periodontal evaluation form algorithm

```
import pandas
df =
pandas.read_csv('PerioTxPlanPresentation_clean.tsv',sep='\t',encodi
ng='latin-1')
new_df = df[['Id','Date',"'a. Diagnosis: '"]].copy()
new_df.to_csv('PerioTxPlanPresentation_Subset.tsv',sep='\t',encodin
g='latin-1')
```

Simultaneously, a *Treatment Subset Extractor.py* algorithm was also created that used IF-ELSE statements to extract four treatments (COE, POE, PM, and PRE) received by the patients from the treatment history master dataset (*e.g. of syntax: if 'D0150' in procedure or 'D0120' in procedure or 'D0127' in procedure or 'D4910'*). When these procedures codes were found in the treatment history master dataset, the Treatment subset extractor.py retrieved this information with patients' ID, procedure completed date, and description of the procedure in a new text file *Treatment_History_Subset. TSV*

Table 44: A computer algorithm to extract comprehensive oral evaluation, periodic oral evaluation, periodontal maintenance, and periodontal re-evaluation information from the treatment history master dataset

```
import pandas as pd
#EXTRACT ALL PROCEDURES
df = pd.read_csv('TrxHistFindings.tsv', sep='\t', encoding='latin-
1')
new_df =
df[['IRB_ID','ModifiedDateTime','Description','Procedure','Site']].
copy()
new_df.to_csv('TrxHistFindings_Subset_ALL_PROCEDURES.tsv',sep='\t',
encoding='latin-1')
#EXTRACT SELECTED PROCEDURES
df = pd.read_csv('TrxHistory.tsv',sep='\t',encoding='latin-1')
output_file = open('TrxHistory_Subset_Selected_Procedures.tsv','w')
output_file.write('IRB_ID\tModifiedDateTime\tDescription\tProcedure
\tSite\n')
for i in range(len(df)):
    procedure = df['Procedure'][i]
    if 'D0150' in procedure or 'D0120' in procedure or 'D0127' in
procedure or 'D4910' in procedure:
        output_file.write('%s\t%s\t%s\t%s\t%s\t\n' %
(df['IRB_ID'][i],df['ModifiedDateTime'][i],df['Description'][i],df[
'Procedure'][i],df['Site'][i]))
output_file.close()
```

Table 45: Converting patients' charting findings such as clinical attachment loss, bleeding on probing, and periodontal pocket depth in individual text files

```
file_name = input("Enter File Name: ")
data = open(file_name,'r').readlines() # Reads all the content
contained in the file and stores it in the 'data' variable
data = list(map(str.strip,data)) # Removes the new line
characters from the 'data' varaiable so program doesn't get
confused
data = list(map(lambda item: item.replace('"',''),data)) #
Replaces all of the double quote characters from the 'data'
variable
has_header = input("Does the file have an identifying header? Y
or N: ")
if has_header.lower() == 'y':
    del(data[0]) # deletes the first row in the file so that the
program doesn't process the identifying header row
for i in range(len(data)): # Runs the code below on every line in
the data
    patient_data = data[i].split(',') # Converts each line into a
list separated by ',' so we can access each individual items in
the line.
    patient_id = patient_data[1]
    date = patient_data[3]
    percond = patient_data[4]
    sitename = patient_data[5]
    section = patient_data[6]
    v1 = patient_data[7]
    v2 = patient_data[8]
    v3 = patient_data[9]
    file_name = ('%s_%s.txt' % (str(patient_id),str(date))) #
Creates a variable 'file_name' which has the format:
patientID_date.txt
    output_file = open(file_name,'a')
    output_file.write('PerCond: %s\nSiteName: %s\nSection:
%s\nValue1: %s\nValue2: %s\nValue3: %s\n' %
(percond,sitename,section,v1,v2,v3)) # Writes all the variables
to the file in separate lines
    output_file.close()
print("Completed!")
```

Table 46: Gingivitis_Diagnoser.py computer algorithm to automatically diagnose patients' gingivitis status into healthy, localized or generalized cases

```
import os,re
def calculate_total_sites(file_content):

    total_sites = 0
    sites_used = [] # list that stores all of the sites that have
been read only once
    for i in range(len(file_content)): # Runs the following code
for every line in the data
        if 'SiteName:' in file_content[i]: # Checks if the
keyword 'SiteName:' exists in the line. If yes: performs the code
below
            current_site = file_content[i][10:] # Creates
variable 'current_site' that stores the sitename eg: 10
            if not current_site in sites_used: # Checks whether
the current_site variable exists in the 'sites_used' list
                sites_used += [current_site]
                total_sites += 1 # Adds 1 to the previous value
of the variable 'total_sites'
    return total_sites * 6

def calculate_bop(file_content):
    """
    This function reads the file and checks for the number of
sites that were bleeding and then returns it.
    """
    bop_sites = 0
    for i in range(len(file_content)):
        if 'PerCond: BLEED' in file_content[i]:
            values =
[file_content[i+3][8:],file_content[i+4][8:],file_content[i+5][8:
]]
            for item in values:
                if item == '1' or item == 'B' or item == 'b':
                    bop_sites += 1

    return bop_sites

def calculate_attach_sites(file_content):
    """
    This function reads the file and checks for the number of
attach sites and returns it
    """
    attach_sites = 0
    for i in range(len(file_content)):
        if 'PerCond: ATTACH' in file_content[i]:
            attach_sites += 1
    return attach_sites
```

```
def diagnose(total_sites, bop_sites,file):
    """
    This function calculates the percentage of the number of
sites that bled and uses the given criteria to move the file into
    its determined diagnosis.
    """
    if total_sites == 0: # Checks if the text file didn't contain
any sites meaning it was an invalid file and moves it to the
Unknown folder
        os.rename(file,'Unknown/'+file)
        return "Unknown"

    elif round((bop_sites / total_sites) * 100) < 10: # Checks if
the bop score is less than 10% if so: moves it to 'No Gingivitis'
folder.
        os.rename(file,'No_Gingivitis/'+file)
        return "No Gingivitis"
    elif round((bop_sites / total_sites) * 100) >= 10 and
round((bop_sites / total_sites) * 100) <= 30: # Checks if the bop
score is less than or equal to 30% and also greater than or equal
to %10 if so: moves it to 'Localized Gingivitis' folder.
        os.rename(file,'Localized_Gingivitis/'+file)
        return "Localized Gingivitis"
    elif round((bop_sites / total_sites) * 100) > 30: # Checks if
the bop score is more than 30% if so: moves it to 'Generalized
Gingivitis' folder.
        os.rename(file,'Generalized_Gingivitis/'+file)
        return "Generalized Gingivitis"

def main():
    files = []
    try: # Tells the program not to crash if it is unable to
create the folders
        os.mkdir('No_Gingivitis')
        os.mkdir('Localized_Gingivitis')
        os.mkdir('Generalized_Gingivitis')
        os.mkdir('Unknown')
    except FileExistsError:
        pass

    for f in os.listdir():
        if re.search('.txt',f):
            files += [f]

    log_file_content = ""
    for file in files: # Performs the following code for every
text file contained in the directory
        data = open(file,'r').readlines() # Reads the data from
the text file and stores it in a variable called 'data'
        data = list(map(str.strip,data)) # Removes the newline
characters from the 'data' variable
```

```
        patient_id = file.split('_')[0] # Gets the first item
contained in the line which is patient id and stores it in a
variable called 'patient_id'
        patient_date = file.split('_')[1].replace('.txt','')
        total_sites = calculate_total_sites(data)
        total_teeth = round(total_sites / 6)
        bop_sites = calculate_bop(data)
        attach_sites = calculate_attach_sites(data)
        if attach_sites == 0: # If the text file is missing
attach information, move the file to the unknown folder.
            log_file_content += ("Patient ID: " + str(patient_id)
+ "\n" + "Date: " + str(patient_date) + "\n" + "Total Teeth:
ATTACH MISSING" + "\n" + "Total Sites: ATTACH MISSING" + "\n" +
"BOP Sites: ATTACH MISSING" + "\n" + "Affected: ATTACH MISSING"+
"\n" + "Diagnosis: UNKNOWN" + "\n\n\n")
            os.rename(file,'Unknown/'+file)
            continue
        diagnosis = diagnose(total_sites,bop_sites,file)

        log_file_content += ("Patient ID: " + str(patient_id) +
"\n" + "Date: " + str(patient_date) + "\n" + "Total Teeth: " +
str(total_teeth) +"\n" + "Total Sites: " + str(total_sites) +
"\n" + "BOP Sites: " + str(bop_sites) + "\n" + "Affected: " +
str(round((bop_sites / total_sites) * 100)) +"%"+ "\n" +
"Diagnosis: " + diagnosis + "\n\n\n")
    log_file = open("log.txt",'w') # Creates a log file
containing each patients' information and the diagnosis
    log_file.write(log_file_content)
    log_file.close()

if __name__=="__main__":
    main()
```

Table 47: A computer algorithm to determine patients' severe periodontitis status using clinical attachment loss and periodontal pocket depth information

```
import os # import is a function and OS is a library. OS library
helps in creating new folders. For example, we are creating a
folder named "severe cases"
import re # regular expression library to find .txt files.

files = [] # files is a list variable in which all the processed
file names 00prw00_05_July_2014.txt are saved and later they will
be moved to either severe or other cases based on the given
criteria below
try:
    os.mkdir("Severe_Cases") # mkdir is a function that creates a
folder named: severe_Cases
    os.mkdir("Others")
except FileExistsError: #if the folder with the same name exists
Severe_Cases, Others then program will not crash but will add
files to those folders.
    pass
for file in os.listdir():
    if re.search('.txt',file): # only consider text files
        files += [file]


# In[28]:


for i in range(len(files)): # Performs the following code in all
of the text files in the directory
    data = open(files[i],"r").readlines() # Opens the first .txt
file and stores all of the data into a variable called 'data'
    data = list(map(str.strip,data))# remove new line characters
from the varaible 'data' so that the program doesn't malfunction
    sites_affected_attach = 0 # A variable that will later be
used to save the total number of teeth affected by a given
criteria
    sites_affected_pocket = 0

    for line in range(len(data)): # Will do the following
instructions for every line
        if 'ATTACH' in data[line]: # in the given print out, if
the first line has the keyword 'ATTACH' then do the following,
and if not, then check next line until 'ATTACH' is found. So in
```

```
this example, first line has ATTACH present therefore, continue
with the instructions below. Here, this line is the 0th line
meaning it is the very first line  (see example document).
            current_sitename = data[line+1][10:] # Because the
above if statement is true, read first line (above described) + 1
(next line). In the example, it will be SiteName: 10 Next, [10:]
represents indexing. Therefore, it will look at the 10th element
of the second line. For example, the second line is "SiteName:
10", here, S is 0, i is 1, t is 2, e is 3, and so on. 1 is the
tenth element and anything after 1. This will pull out tooth
number "10".
            try:
                next_sitename = data[line+7][10:] # look at the
line + 7, and save the tooth number in the variable next_sitename
            except IndexError: # if the file ends then stop the
whole thing
                break

            if current_sitename == next_sitename: # if both of
the site names are similar, in this case, "10", then do the
following. And if not, then read the next line
                all_values = [] # all_values is a list variable
that will stores all of the values named Value1,Value2,Value3 for
both sites 1 and 0
                try:
                    all_values += [int(data[line+3][8:])] # look
at the line 3, and 8th element which is the value mm of
attachment and convert into integer (5).
                except ValueError: # if missing value then
consider 0. We found that number of patients had either facial or
lingual sites missing, and some patients had at least one of six
sites missing.
                    all_values += [0] #if it doesn't work then
add 0 (if missing)
                try:
                    all_values += [int(data[line+4][8:])] # look
at the line 4, and 8th element which is the value (mm) of
attachment and convert into integer (3).
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # look
at the line 5, and 8th element which is the value (mm) of
attachment and convert into integer (5).
                except ValueError:
```

```
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # look
at the line 9, and 8th element which is the value (mm) of
attachment and convert into integer (4).
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # look
at the line 10, and 8th element which is the value (mm) of
attachment and convert into integer (4).
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+11][8:])] # look
at the line 11, and 8th element which is the value (mm) of
attachment and convert into integer (4).
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 6: # if it finds any value
>= 6 from the variable 'all_values' then it adds 1 to the
sites_affected_attach varaible.
                    sites_affected_attach += 1 # previous value
of sites_affected_attach + 1
                all_values = [] # Empties the all_values variable
so next set of values can be read
            else:
                continue


    for line in range(len(data)): # Will do the following
instructions for every line
        if 'POCKET' in data[line]: # in the given print out, if
the first line has PerCond=POCKET then do the following, and if
not, then check next line unitl PerCond=POCKET found.
            current_sitename = data[line+1][10:] # Because the
above if statement is true, read first line (above described) + 1
(next line).
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
```

```
                try:
                    all_values += [int(data[line+3][8:])] # look
at the line 3
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] #  look
at the line 4
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] #  look
at the line 5
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] #  look
at the line 9
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] #
look at the line 10
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+11][8:])] #
look at the line 11
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 5:
                    sites_affected_pocket += 1
                all_values = []
            else:
                continue


    if sites_affected_attach >= 2 and sites_affected_pocket >= 1:
# If the total number of ATTACH sites that met the criteria is
greater than or equal to 2 AND the total number of POCKET sites
that met the criteria is greater than or equal to 1: Then move
the file to Severe_Cases folder
        os.rename(str(files[i]),"Severe_Cases/"+str(files[i]))
    else:
```

```
        os.rename(str(files[i]),"Others/"+str(files[i])) # If not
then move the file to the Others folder
```

Table 48: A computer algorithm to determine patients' moderate periodontitis status based on clinical attachment loss and periodontal pocket depth information

```
#!/usr/bin/env python
# coding: utf-8

# In[29]:


import os
import re


# In[27]:


files = []
try:
    os.mkdir("Moderate_Cases")
    os.mkdir("Others")
except FileExistsError:
    pass
for file in os.listdir():
    if re.search('.txt',file):
        files += [file]


# In[28]:


for i in range(len(files)):
    data = open(files[i],"r").readlines()
    data = list(map(str.strip,data))
    sites_affected_attach = 0
    sites_affected_pocket = 0

    for line in range(len(data)):
        if 'ATTACH' in data[line]:
            current_sitename = data[line+1][10:]
```

```
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
                try:
                    all_values += [int(data[line+3][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+11][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 4:
                    sites_affected_attach += 1
                all_values = []
            else:
                continue


    for line in range(len(data)):
```

```
        if 'POCKET' in data[line]:
            current_sitename = data[line+1][10:]
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
                try:
                    all_values += [int(data[line+3][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+11][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 5:
                    sites_affected_pocket += 1
                all_values = []
            else:
                continue
```

```
    if sites_affected_attach >= 2 or sites_affected_pocket >= 2:
        os.rename(str(files[i]),"Moderate_Cases/"+str(files[i]))
    else:
        os.rename(str(files[i]),"Others/"+str(files[i]))
```

Table 49: A computer algorithm to determine patients' mild periodontitis status based on clinical attachment loss, and periodontal pocket depth information

```
#!/usr/bin/env python
# coding: utf-8

# In[29]:


import os
import re


# In[27]:


files = []
try:
    os.mkdir("Mild_Cases")
    os.mkdir("Others")
except FileExistsError:
    pass
for file in os.listdir():
    if re.search('.txt',file):
        files += [file]


# In[28]:


for i in range(len(files)):
```

```
    data = open(files[i],"r").readlines()
    data = list(map(str.strip,data))
    sites_affected_attach = 0
    sites_affected_pocket = 0
    sites_affected_pocket2 = 0

    for line in range(len(data)):
        if 'ATTACH' in data[line]:
            current_sitename = data[line+1][10:]
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
                try:
                    all_values += [int(data[line+3][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+11][8:])] # Adds
value1
                except ValueError:
```

```
                    all_values += [0]
                if max(all_values) >= 3:
                    sites_affected_attach += 1
                all_values = []
            else:
                continue


    for line in range(len(data)):
        if 'POCKET' in data[line]:
            current_sitename = data[line+1][10:]
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
                try:
                    all_values += [int(data[line+3][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
```

```
                    all_values += [int(data[line+11][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 4:
                    sites_affected_pocket += 1
                all_values = []
            else:
                continue

    for line in range(len(data)):
        if 'POCKET' in data[line]:
            current_sitename = data[line+1][10:]
            try:
                next_sitename = data[line+7][10:]
            except IndexError:
                break

            if current_sitename == next_sitename:
                all_values = []
                try:
                    all_values += [int(data[line+3][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+4][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+5][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+9][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                try:
                    all_values += [int(data[line+10][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
```

```
                try:
                    all_values += [int(data[line+11][8:])] # Adds
value1
                except ValueError:
                    all_values += [0]
                if max(all_values) >= 5:
                    sites_affected_pocket2 += 1
                all_values = []
            else:
                continue


    if (sites_affected_attach >= 2 and sites_affected_pocket >=
2) or (sites_affected_pocket2 >= 1):
        os.rename(str(files[i]),"Mild_Cases/"+str(files[i]))
    else:
        os.rename(str(files[i]),"Others/"+str(files[i]))
```

Table 50: Text-mining program to extract patients' clinician-recorded diagnosis from periodontal evaluation forms

```
Merge files: Demographics and clinician-recorded diagnosis
"""
import pandas as pd
from collections import defaultdict



df1 = pd.read_excel('NEW_PERIOTX_07302019.xlsx',sheet_name=0)
df2 = pd.read_excel('NEW_PERIOTX_07302019.xlsx',sheet_name=1)

dic={}
l_birth=defaultdict(list)
l_sex=defaultdict(list)
l_race=defaultdict(list)
l_insurance=defaultdict(list)


'''
l_id=df1['Id'].tolist()
for i,pid in enumerate(l_id):
```

```
    if pid in df2['id']:
        l_birth.append(df2.iloc[
                1,i])#1 beacuse birth is column 1
        l_sex.append(df2.iloc[2,i])
        l_race.append(df2.iloc[3,i])
        l_insurance.append(df2.iloc[4,i])
'''

df3=pd.merge(df1,df2,how='left')
print(df3)

writer = pd.ExcelWriter('try.xlsx')
df3.to_excel(writer)
writer.save()
```

Text mining program to convert clinician-recorded diagnosis in structured format:

```
# -*- coding: utf-8 -*-
"""
Created on Mon Jul 16 08:43:07 2018

@author: kumarkr
"""

import nltk
import string
import re
import pandas as pd

global_loc=[]
global_reg=[]
global_disease=[]
global_time=[]
global_severity=[]
global_id=[]
global_diagnosis=[]
global_date=[]
global_birth=[]
global_race=[]
global_sex=[]
global_insurance=[]


def get_time_period(words):
```

```
    tp=''
    labels = ['chronic', 'acute']
    dic = {}
    for l in labels:
        dic[l] = 0
    for w in words:
        for l in labels:
            ed = nltk.edit_distance(l, w)
            if ed < 3:
               # print('ed ={}'.format(l))
                dic[l] += 1
    if dic['chronic'] >= 1:
        tp= 'chronic'
    elif dic['acute'] >= 1:
        tp = 'acute'
    else:
        tp='No time period'
    return tp

def get_disease(words):
    disease = ''
    labels = ['gingivitis','periodontitis']
    dic = {}
    for l in labels:
        dic[l] = 0
    for w in words:
        for l in labels:
            ed = nltk.edit_distance(l, w)
            if ed < 5:
                #print('ed ={}'.format(l))
                dic[l] += 1
    if dic['gingivitis']>=1:
        disease='gingivitis'
    elif dic['periodontitis']>=1:
        disease='periodontitis'
    else:
        disease='No disease specified'

    return disease

def get_severity(words):
    severity=''
    labels = ['mild', 'moderate', 'severe']
    dic = {}
    for l in labels:
```

```
        dic[l]=0
    for w in words:
        for l in labels:
            ed=nltk.edit_distance(l,w)
            if ed<3:
                # print(w)
                # print('ed ={}'.format(l))
                dic[l]+=1
    if dic['mild']>=1 and dic['moderate']>=1:
        severity='mild to moderate'
    elif dic['moderate']>=1 and dic['severe']>=1:
        severity='moderate to severe'
    elif dic['mild']==1:
        severity='mild'
    elif dic['moderate']==1:
        severity='moderate'
    elif dic['severe']==1:
        severity="severe"
    else:
        severity = 'no abnormality'

    return severity


def get_reg(words):
    region_labels = ['maxillary', 'mandibular']
    for w in words:
        for l in region_labels:
            ed =nltk.edit_distance(l, w)
            if ed < 4:
                return l
    teeth=[]
    for w in words:
        '''if '-' in w or '&' in w:
            if '-' in w:
                s_range=w.split('-')
            else:
                s_range=w.split('&')
            if is_number(s_range[0]) and is_number(s_range[1]):
                for i in range(int(s_range[0]),int(s_range[1])+1):
                    teeth.append(i)'''
        if hasNumbers(w):
            teeth.append(w)
    if len(teeth)>0:
        return teeth
```

```
    return 'No region specified'

def get_loc_pos(words):
    loc_labels = ['generalized', 'localized']
    location_list = []
    for i, w in enumerate(words):
        for l in loc_labels:
            ed =nltk.edit_distance(l, w)
            if ed < 4:
               location_list.append(i)
    return location_list

def get_loc(words):
    loc_labels = ['generalized','localized']
    location_tuples=[]
    for i,w in enumerate(words):
        for l in loc_labels:
            ed = nltk.edit_distance(l, w)
            if ed < 4:
                return l
    return 'No location specified'

def isNan(x):
    return x!=x


def hasNumbers(inputString):
    return any(char.isdigit() for char in inputString)

def is_number(s):
    try:
        float(s)
        return True
    except ValueError:
        return False

def read_xcel():
    df = pd.read_excel('try.xlsx')
    print(df.columns)
    #print(df['Id'])
    for i,s in enumerate(df['diagnosis']):
        if isNan(s)==False:
            #print(df['Id'][df.index[i]])
```

```
clean(s,df['Id'][df.index[i]],df['Date'][df.index[i]],df['birth']
[df.index[i]],df['sex'][df.index[i]],df['race'][df.index[i]],df['
Insurance'][df.index[i]])



def clean(diag,pid,date,birth,sex,race,insurance):
   words = re.split(r'[^a-zA-Z0-9_&]', diag)
   pat=re.compile('\d+')
   numbers=pat.findall(diag)
   # table = str.maketrans('', '', string.punctuation)
   # words = [w.translate(table) for w in words]
   words=[word.lower() for word in words]
   #stop_words = set(stopwords.words('english'))
   stop_words={'needn', 'mightn', 'a', 'not', 'then', 'ours',
'wouldn', 'those', 'our', "doesn't", 'having', 'again', 'most',
               'mustn', 'his', 'd', 'below', 'when', 'only',
"isn't", 've', "mightn't", 'during', "you'll", 'is', 'can',
'couldn', "wasn't",
               'were', 'at', 'both', 'by', 'other', 'about',
"you're", 'some', 'ain', 'your', 'yours', 'hasn', 'until',
'above',
               'you', 'very', 'few', 'herself', 'they', 'on',
"mustn't", 'why', 'didn', 'no', "needn't", 'themselves',
'should',
               'shouldn', 'aren', 'don', 'shan', 'himself', 'or',
'it', 'as', 'so', 'did', 'she', 'and', 'hers', 'ma', 'll',
"won't",
               "should've", 'where', 'the', 'over', "don't",
'who', 'off', 'we', 'all', 'if', 's', 'its', 'any', 'than', 'me',
"weren't",
               'am', 'do', 'there', 'here', 'which', "you'd",
'because', 'was', 'weren', 'being', 'be', 'further', 'each',
'whom', 'her',
               'out', "shan't", 'an', 'haven', 'in', 'been',
'under', 'same', 'o', 'theirs', "that'll", "you've", 'but', 'y',
'won', 'i',
               'to', 'nor', 'them', 'against', 'for', "couldn't",
'yourselves', 'these', "hasn't", 'now', 't', 'own', 'between',
'up', "it's",
               're', 'has', 'doesn', 'how', 'such', "wouldn't",
"didn't", 'through', "haven't", 'isn', "shouldn't", 'he', 'my',
'him', "hadn't",
```

```
                'yourself', 'while', 'will', 'ourselves', 'their',
'itself', 'what', 'does', 'are', 'have', 'into', 'wasn', 'of',
'hadn', "she's",
                'doing', 'after', 'once', 'm', 'this', 'had',
'more', 'that', 'just', 'down', "aren't", 'with', 'from',
'before', 'myself', 'too'}
   words = [w for w in words if not w in stop_words and len(w)>0]
   loc_list=get_loc_pos(words)
   last_end=0
   for start,end in zip(loc_list,loc_list[1:]):
      location=words[start]

my_print(words[start:end+1],diag,pid,date,birth,sex,race,insuranc
e)
      last_end=end

my_print(words[last_end:],diag,pid,date,birth,sex,race,insurance)

def my_print(words,diagnosis,pid,date,birth,sex,race,insurance):
   disease = get_disease(words)
   severity = get_severity(words)
   time_period = get_time_period(words)
   reg=get_reg(words)

   location=(get_loc(words))
   print('location={}'.format(location))
   print('region={}'.format(reg))
   print('disease={}'.format(disease))
   print('severity={}'.format(severity))
   print('time period ={}'.format(time_period))


   global global_loc
   global global_reg
   global global_disease
   global global_time
   global global_severity
   global global_id
   global global_diagnosis
   global global_birth
   global global_sex
   global global_race
   global global_insurance
   global global_date
```

```
    global_loc.append(location)
    global_reg.append(reg)
    global_disease.append(disease)
    global_time.append(time_period)
    global_severity.append(severity)
    global_id.append(pid)
    global_diagnosis.append(diagnosis)
    global_birth.append(birth)
    global_sex.append(sex)
    global_race.append(race)
    global_insurance.append(insurance)
    global_date.append(date)


def main():
    writer = pd.ExcelWriter('cleaned_diagnosis_small.xlsx')
    read_xcel()

n_df=pd.DataFrame({'pid':global_id,'diagnosis':global_diagnosis,'
disease':global_disease,'time':global_time,'severity':global_seve
rity,'location':global_loc,'region':global_reg,'birth':global_bir
th,'sex':global_sex,'race':global_race,'insurance':global_insuran
ce,'date':global_date})
    #print(n_df)
    n_df.to_excel(writer)
    writer.save()

if __name__ == '__main__':
    main()


import nltk
import string
import re
import pandas as pd
from collections import defaultdict

df = pd.read_excel('structured_diagnosis_final.xlsx')
d=defaultdict(list)
time=defaultdict(list)
for i,v in enumerate(df['pid']):
    s=str(df.iloc[i]['location'])+'
'+str(df.iloc[i]['severity'])+' '+str(df.iloc[i]['disease'])+'
'+str(df.iloc[i]['time'])
    d[v].append(s)
```

```
    date=str(df.iloc[i]['date'])
    d[v].append(date)

writer = pd.ExcelWriter('output_progression_structured.xlsx')

n_df=pd.DataFrame(list(d.items()),columns=['pid','td'])
n_df2=pd.DataFrame(n_df.td.values.tolist())
n_df2['pid']=list(d.keys())




df=df.drop(labels=['diagnosis','disease','location','region','sev
erity','time','date'],axis=1)
#
df=df.drop_duplicates(subset='pid')
df=df.set_index('pid')
n_df2=n_df2.join(df,on='pid')



n_df2.to_excel(writer)
writer.save()

Putting clinician-recorded diagnoses in buckets.

# -*- coding: utf-8 -*-
"""
Created on Tue Jul 17 08:27:11 2018

@author: kumarkr
"""

import nltk
import string
import re
import pandas as pd
from collections import defaultdict

df = pd.read_excel('cleaned_diagnosis_small.xlsx')
d=defaultdict(list)
time=defaultdict(list)
for i,v in enumerate(df['pid']):
```

```
    s=str(df.iloc[i]['location'])+'
'+str(df.iloc[i]['severity'])+' '+str(df.iloc[i]['disease'])+'
'+str(df.iloc[i]['time'])
    d[v].append(s)
    date=str(df.iloc[i]['date'])
    d[v].append(date)

writer = pd.ExcelWriter('bin_diagnosis_small.xlsx')

n_df=pd.DataFrame(list(d.items()),columns=['pid','td'])
n_df2=pd.DataFrame(n_df.td.values.tolist())
n_df2['pid']=list(d.keys())



df=df.drop(labels=['diagnosis','disease','location','region','sev
erity','time','date'],axis=1)
#
df=df.drop_duplicates(subset='pid')
df=df.set_index('pid')
n_df2=n_df2.join(df,on='pid')



n_df2.to_excel(writer)
writer.save()
```

Table 51: Text-mining program to generate number of patients' whose PD severity changed over time between their first and last visits

```
import nltk
import string
import re
import pandas as pd
from collections import defaultdict
import re

def isNan(x):
```

```
    return x!=x


df = pd.read_excel('bin_diagnosis_small_NEW.xlsx')


dic_change=defaultdict(int)
severity_list=['mild','mild to moderate','(?<!to
)moderate','moderate to severe','(?<!to )severe']
#severity_list=['mild','mild to moderate']

disease_list=['gingivitis','periodontitis','No disease']
location_list=['localized','generalized','No location specified']
#print(df.head())
#print(df.count(axis=1).values)
#row_lens=df.count(axis=1).values
#print(row_lens[12])
#print(row_lens)
#print(type(row_lens))
count=0
severity=''
disease=''
location=''
o_disease=''
o_location=''
o_severity=''
res_df=pd.DataFrame(columns=['pid','from','to'])
from_list=[]
to_list=[]
pid_list=[]

for i ,row in df.iterrows():
    for n,cell_vals in enumerate(row):
        try:
            if(isNan(cell_vals)):
                severity=''
                disease=''
                location=''
                o_disease=''
                o_location=''
                o_severity=''
                break
            if n%2==1 and n<16:
                #print(n)
                if len(disease)>0:
```

```
                    if location =='localized' or
location=="generalized":
                        o_disease=disease
                        o_location=location
                        o_severity=severity

                # print(cell_vals)
                # print(type(cell_vals))
                for s in severity_list:
                    v=re.search(s,cell_vals)
                    if v:
                        severity=v.group()
                        #print(severity)
                for d in disease_list:
                    v=re.search(d,cell_vals)
                    if v:
                        disease=v.group()
                        #print(disease)
                for l in location_list:
                    v=re.search(l,cell_vals)
                    if v:
                        location=v.group()
                        #print(location)
                if len(o_disease)>0:
                    if str(row[n+1])!=str(row[n-1]):
                            #print(n)
                        print(type(row))
                        print(row[n+1])
                        print(row[n-1])
                        temp1=o_location+' '+o_severity+'
'+o_disease+' '
                        temp2=location+' '+severity+' '+disease+'
'
                        pid_list.append(row['pid'])
                        tup=(temp1,temp2)
                        print(tup)
                        from_list.append(temp1)
                        to_list.append(temp2)
                        count+=100
        except :
            continue


res_df['from']=from_list
res_df['to']=to_list
```

```
res_df['pid']=pid_list
#res_df['pid']=row['pid']
dic=defaultdict(int)
for f,t in zip(from_list,to_list):
    tup=(f,t)
    dic[tup]+=1
n_df=pd.DataFrame.from_dict(dic,orient='index')

# print(n_df)
# #print(res_df.tail())
# print(count)

writer1 = pd.ExcelWriter('diagnosis_pot_counter2_NEW.xlsx')
writer2=pd.ExcelWriter('diagnosis_pot_bins2_NEW.xlsx')

n_df.to_excel(writer2,sheet_name='Sheet1')
res_df.to_excel(writer1,sheet_name='Sheet1')

writer1.save()
writer2.save()
# print(df.dtypes)
```

Table 52: A computer algorithm to extract patient information from their first comprehensive oral examination

```
import pandas as pd
import numpy as np

df_diabetes = pd.read_excel('ALL_Cases.xlsx',sheet_name=0)
df_procedures = pd.read_excel('Procedures.xlsx',sheet_name=0)
df_procedures['Offset Date'] =
df_procedures['ModifiedDateTime'].apply(lambda x: x-
pd.DateOffset(years=1))

def get_index(patient_id):
    return df_procedures.index[df_procedures['IRB_ID'] ==
patient_id].tolist()

    for i in range(len(df_diabetes)):
    found = False
```

```
    patient_id = df_diabetes['Patient ID'][i]
    date = df_diabetes['Date'][i]
    pos_in_procedures = get_index(patient_id)

    for position in pos_in_procedures:
        offset_date = df_procedures['Offset Date'][position]
        if (('D0150' in df_procedures['Procedure'][position])
and (date >= offset_date and date <=
df_procedures['ModifiedDateTime'][position])):
            found = True
            df_diabetes['Procedure Code Date'][i] =
df_procedures['ModifiedDateTime'][position]
            df_diabetes['Found'][i] = 'TRUE'
            break
        else:
            found = False
            continue
    if found == False:
        df_diabetes['Found'][i] = 'FALSE'
    print("Entry %s completed." % i)
```

Manual review guidelines to evaluate performance of the gingivitis diagnoser.py program.

Objective

The objective of this manual review process is to evaluate the **performance of gingivitis diagnoser.py program** that automatically generated gingivitis diagnoses from periodontal charting findings.

PERIODONTAL CHARTING

Each file manually reviewed by the experts had the below information:

- patient ID
- appointment date
- diagnosis generated automatically by gingivitis diagnose. py program based on charting findings
- Complete periodontal charting

Manual Review Process

First, experts manually reviewed 50 common patients' full charting information and manually diagnose their gingivitis status based on the criteria described in Table 6. The inter-rater agreement between the faculties was 0.9 (Cohen's Kappa value) which indicated excellent agreement. Next, each expert reviewed 150 records independently, which resulted in an overall dataset of 350 cases. Next, the diagnoses automatically generated by gingivitis_diagnoser.py was compared with experts' diagnoses. Based on the computer algorithm's ability to correctly diagnose gingivitis cases, true positives, false positives, and false negatives were calculated. Using these measures, the performance of

gingivitis_diagnoser.py was determined. Last, precision, recall, and f-measure were calculated using the formulas described in Table 7.

Table 53: Diagnosis Criteria for gingivitis [126]

Disease status	Rules
No gingivitis	No presence of "B" or "1" correspond to "BLEED" value in the charting text file OR **the BOP score is <10%.**
Localized gingivitis	When the BOP score is >= **10% AND <=30%.**
Generalized gingivitis	When BOP score is **more than 30%.**

Table 54: Formulas to evaluate the performance of gingivitis_diagnoser.py on testing dataset

Evaluation measures	Formulas
Precision	true positive / (true positives + false positives)
Recall	true positives / (true positive + false negatives)
F-measure	2 * (precision * recall) / (precision + recall)

For the manual review, a gingivitis validation calculator in Excel was calculated. Experts first counted the total number of bleeding sites manually in the charting file and manually entered this information in the BOP sites. If there are no bleeding sites, then they considered bleeding sites = 0. Next, they calculated total number of teeth by manually reviewing the number of teeth that had CAL information recorded. Next, they determined the BOP score by dividing the total number of bleeding sites by total number of sites (total teeth * 6 probing sites).

Manual review guidelines to assess the performance of Natural Language Processing Program.

Objective

Determine the performance of a natural language processing (NLP) algorithm that automatically extract patients' clinician-recorded diagnosis in a structured format.

Instructions to manual review process

- To determine NLP algorithm's performance, first, reviewers compared the output generated by the NLP program with the clinician-recorded diagnoses. As described in the manuscript (see manuscript Page 5), clinicians' document patients' following information.
    - PD type: Gingivitis or periodontitis.
    - Disease severity: Mild, mild to moderate, moderate, moderate to severe, severe.
    - Disease onset: Acute or chronic.
    - Disease extent: Localized or generalized.
    - Disease location: Maxilla, mandible, tooth #, etc.

  For example, "chronic generalized mild to moderate periodontitis"

  Therefore, the NLP program classified a patient's PD diagnosis based on PD type, severity, onset, extent and location.
- Next, based on the output generated by the NLP program, reviewers classified the program's output into true positive, false positive, and false negative categories. Steps for categorizing an NLP program's output in true positive, false positive, and false negative are described below.
    - **True positive:** A true positive is an outcome where the algorithm correctly identifies the positive disease status.

      For example, if the clinician-recorded diagnosis stated, **"mild chronic generalized periodontitis"**. The natural language processing algorithm has extracted this diagnosis as 1) **disease type=periodontitis**, 2)

severity=mild, 3) **region=generalized**, and **onset=chronic**, then this class is assigned as "true positive" class (See example Table 1 below).

Table 55: Example table that shows how to determine the true positive categories for a clinical recorded diagnosis of "**Mild chronic generalized periodontitis"**

NLP Diagnosis	Class	severity	Onset	Class	Extension	Class	Region	Class
Periodontitis	**TP**	Mild	Chronic	**TP**	Generalized	**TP**	NS	**TP**

- **False positive:** A false positive is an outcome where the algorithm categorizes a positive disease status when the patient does not have the disease. For example, if NLP algorithm categorizes a patient's "no disease/healthy" status into "gingivitis or periodontitis" then it's a false positive case (See example Tables 2, 3 below).

Table 56: Example table that shows how to determine the false positive categories for a clinical recorded diagnosis of "**Mild acute generalized gingivitis"**

NLP Diagnosis	Class	severity	Onset	Class	Extension	Class	Region	Class
Periodontitis	**FP**	Mild	**Chronic**	**FP**	Generalized	**TP**	NS	**TP**

Table 57: Possible examples to correctly identify false positive cases

Actual disease status (clinician recorded diagnosis)	NLP Algorithm's Output	False Positives
Disease Status		
Healthy	Gingivitis	False Positive
Healthy	Periodontitis	False Positive
Gingivitis	Periodontitis	False Positive
Severity		
Mild	Moderate	False Positive
Mild	Severe	False Positive
Moderate	Severe	False Positive
Onset		
Acute	Chronic	False Positive
Extent		
Localized	Generalized	False Positive

- **False negative:** A false negative is an outcome where the algorithm classifies a patient's disease status as 'no disease or healthy' although the patient has gingivitis or periodontitis. For example, if NLP algorithm classifies a patient who have a "gingivitis or periodontitis" diagnosis in the clinical notes as "no disease/healthy" then it's considered false negative case (See example Tables 4, 5 below).

Table 58: Example table that shows how to determine the false negative categories for a clinical recorded diagnosis of "Mild chronic generalized gingivitis"

NLP Diagnosis	Class	severity	Onset	Class	Extension	Class	Region	Class
No disease	**FN**	Mild	**chronic**	**TP**	**Localized**	**FN**	NS	**TP**

Table 59: Possible examples to correctly identify false negative cases

Actual disease status (clinician recorded diagnosis)	**NLP Algorithm's Output**	**False Negative**
Disease Status		
Gingivitis	Healthy	False Negative
Periodontitis	Healthy	False Negative
Severity		
Moderate	Mild	False Negative
Severe	Mild	False Negative
Severe	Moderate	False Negative
Onset		
Chronic	Acute	False Negative
Extension		
Generalized	Localized	False Negative

To determine true positive, false positive, and false negative cases, reviewers compared disease type, onset, severity, location, K extent with clinician-recorded diagnoses. Based on the program's identification and performance, they entered true positive, false positive, and false negatives.

Manual review guidelines to determine the reasons for the agreements and disagreements between findings generated diagnoses and clinician-recorded diagnoses.

Objective

The objective of this manual review process was to determine the **reasons for the disagreements between diagnoses** generated from findings and clinician-recorded diagnoses.

Manual review process

Experts reviewed patients' following information and diagnose patients' PD status (gingivitis and periodontitis) using their experience and expertise. Dr. Dan Shin, one of the reviewers, is a periodontist and an assistant professor at the department of periodontology at IUSD. Dr. Lisa Willis, second reviewer, is a dentist and a clinical assistant professor at the department of Cariology, Operative Dentistry, and Dental Public Health at IUSD.

- patient ID
- appointment date
- diagnosis generated automatically by algorithms based on charting findings
- Description of the condition: patient's gingival information (gingival color, contour, tippling, plaque index, and calculus index)
- Extent, and pattern: patient's radiographic findings (vertical, horizontal bone loss, location of bone loss, and amount of bone loss)
- etiology and predisposing factors

Based on the detailed information on gingival health and bone loss, please diagnose patients' periodontal health in the "diagnosis" and "reason for diagnosis" sections. For reference, experts used the classification described in Table 3 that was used by the dental clinicians at IUSD between Jan 1, 2009 to Dec 31, 2014.

Table 60: Guidelines to diagnose severity of Periodontitis developed by American Academy of Periodontology [32]

Periodontal Findings	Slight (Mild)	Moderate	Severe (Advanced)
Probing depths	> 3 & < 5 mm	>= 5 & < 7 mm	>= 7 mm
Bleeding on probing	Yes	Yes	Yes
Radiographic bone loss	Up to 15% of root length or >= 2 mm & <=3 mm	16% to 30% or > 3 mm & <=5 mm	> 30% or > 5 mm
Clinical attachment loss	1 to 2 mm	3 to 4 mm	>= 5 mm

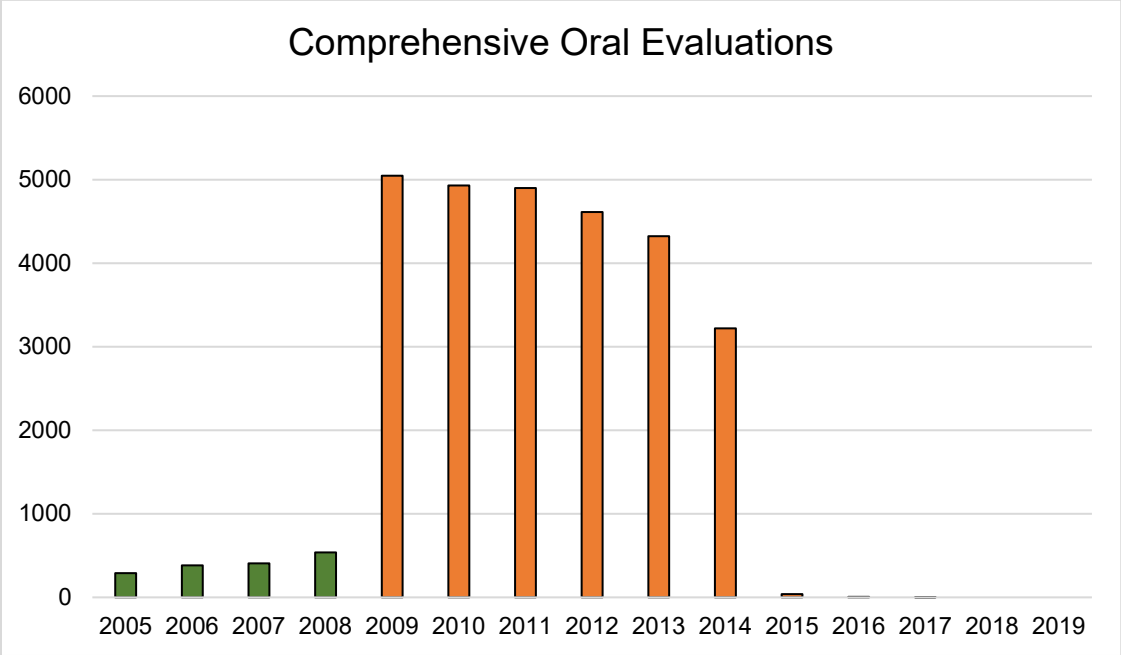


Figure 17: Patients' first comprehensive oral evaluation who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

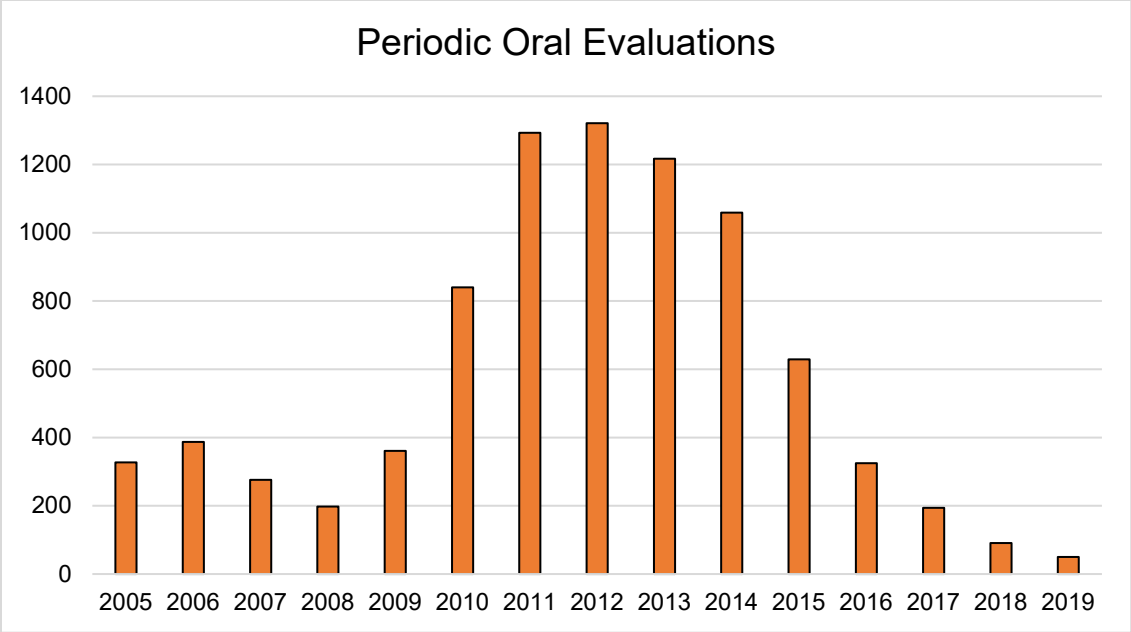


Figure 18: Patients' first periodic oral evaluation who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

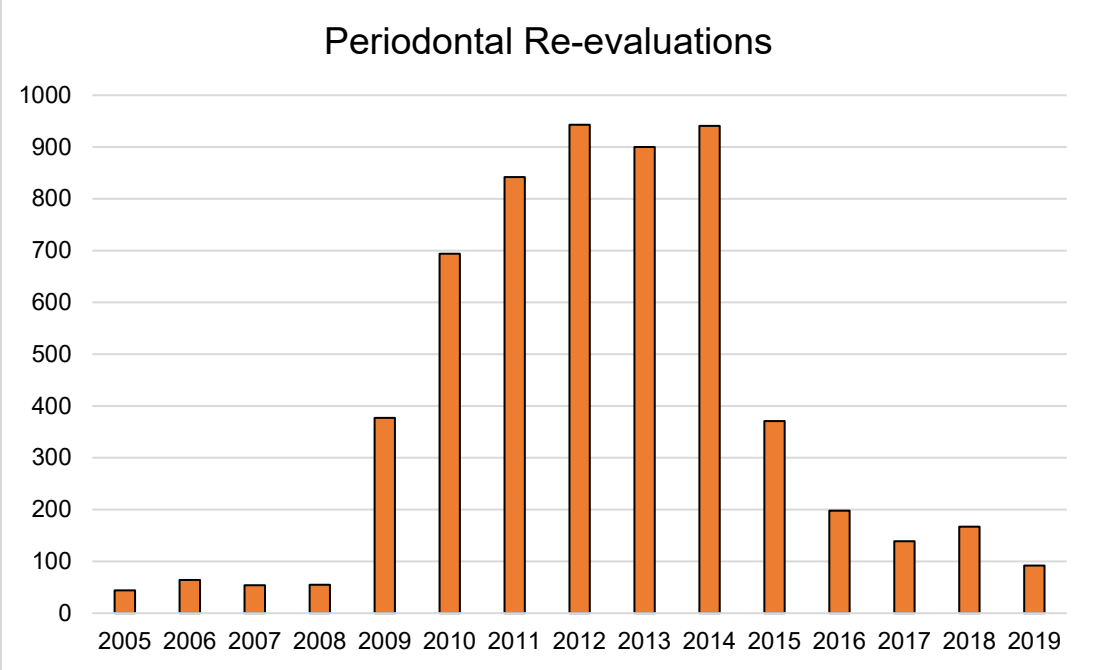


Figure 19: Patients' first periodic oral evaluation who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

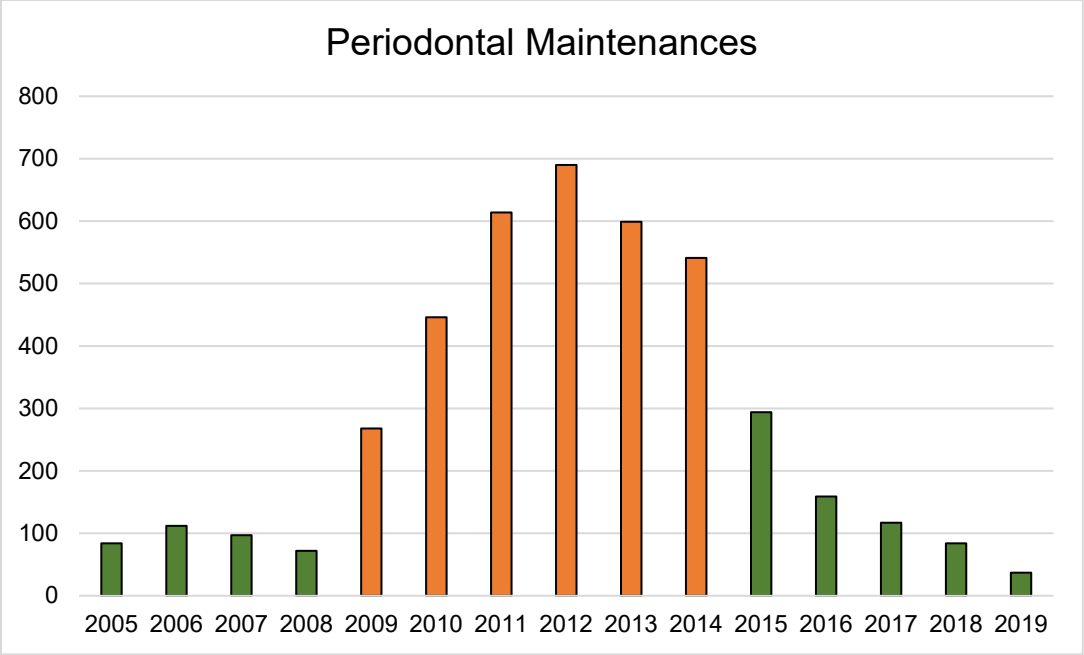


Figure 20: Patients' first periodontal maintenance who received at least one comprehensive oral evaluation between January 1, 2009 to December 31, 2014

10: Bibliography

1. Acharya, A., VanWormer, J. J., Waring, S. C., Miller, A. W., Fuehrer, J. T., & Nycz, G. R. (2013). Regional epidemiologic assessment of prevalent periodontitis using an electronic health record system. *American journal of epidemiology, 177*, 700-707.

2. Albandar, J. M. (1990). A 6-year study on the pattern of periodontal disease progression. *J Clin Periodontol, 17*(7 Pt 1), 467-471. doi:10.1111/j.1600-051x.1990.tb02346.x

3. Albandar, J. M., Baghdady, V. S., & Ghose, L. J. (1991). Periodontal disease progression in teenagers with no preventive dental care provisions. *J Clin Periodontol, 18*(5), 300-304. doi:10.1111/j.1600-051x.1991.tb00432.x

4. AlJehani, Y. A. (2014). Risk factors of periodontal disease: review of the literature. *International journal of dentistry, 2014*.

5. Alwhaibi, M., Balkhi, B., Alshammari, T. M., AlQahtani, N., Mahmoud, M. A., Almetwazi, M., .Alhawassi, T. (2019). Measuring the quality and completeness of medication-related information derived from hospital electronic health records database. *Saudi Pharmaceutical Journal, 27*, 502-506. doi:10.1016/j.jsps.2019.01.013

6. American Academy of Periodontology Task Force Report on the Update to the 1999 Classification of Periodontal Diseases and Conditions. (2015). *J Periodontol, 86*(7), 835-838. doi:10.1902/jop.2015.157001

7. Anderson, A. E., Kerr, W. T., Thames, A., Li, T., Xiao, J., & Cohen, M. S. (2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *J Biomed Inform, 60*, 162-168. doi:10.1016/j.jbi.2015.12.006

8. Armitage, G. C. (1995). Clinical evaluation of periodontal diseases. *Periodontology 2000, 7*, 39-53. doi:10.1111/j.1600-0757.1995.tb00035.x

9. Periodontal diseases: diagnosis., 1 37-215 (1996).

10. Armitage, G. C. (1999). Development of a classification system for periodontal diseases and conditions. *Annals of periodontology, 4*, 1-6.

11. Axelsson, P., & Lindhe, J. (1981). The significance of maintenance care in the treatment of periodontal disease. *J Clin Periodontol, 8*(4), 281-294. doi:10.1111/j.1600-051x.1981.tb02039.x

12. Barbosa, V. L., Angst, P. D., Finger Stadler, A., Oppermann, R. V., & Gomes, S. C. (2016). Clinical attachment loss: estimation by direct and indirect methods. *Int Dent J, 66*(3), 144-149. doi:10.1111/idj.12218

13. Bouchard, P., Carra, M. C., Boillot, A., Mora, F., & Rangé, H. (2017). Risk factors in periodontology: a conceptual framework. *Journal of clinical periodontology, 44*, 125-131.

14. Bruland, P., McGilchrist, M., Zapletal, E., Acosta, D., Proeve, J., Askin, S., .Dugas, M. (2016). Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol, 16*(1), 159. doi:10.1186/s12874-016-0259-3

15. Burcham, W. K., Romito, L. M., Moser, E. A., & Gitter, B. D. (2019). Analyzing Medication Documentation in Electronic Health Records: Dental Students' Self-Reported Behaviors and Charting Practices. *J Dent Educ, 83*(6), 687-696. doi:10.21815/jde.019.070

16. Callahan, T., Barnard, J., Helmkamp, L., Maertens, J., & Kahn, M. (2017). Reporting Data Quality Assessment Results: Identifying Individual and Organizational Barriers and Solutions. *EGEMS (Wash DC), 5*(1), 16. doi:10.5334/egems.214

17. Cekici, A., Kantarci, A., Hasturk, H., & Van Dyke, T. E. (2014). Inflammatory and immune pathways in the pathogenesis of periodontal disease. *Periodontology 2000, 64*, 57-80.

18. Chambrone, L., Chambrone, D., Lima, L. A., & Chambrone, L. A. (2010). Predictors of tooth loss during long-term periodontal maintenance: a systematic review of observational studies. *J Clin Periodontol, 37*(7), 675-684. doi:10.1111/j.1600-051X.2010.01587.x

19. Chan, C. L., You, H. J., Lian, H. J., & Huang, C. H. (2016). Patients receiving comprehensive periodontal treatment have better clinical outcomes than patients receiving conventional periodontal treatment. *Journal of the Formosan Medical Association, 115*, 152-162. doi:10.1016/j.jfma.2015.10.017

20. Chatzopoulos, G. S., Cisneros, A., Sanchez, M., & Wolff, L. F. (2018). Systemic medical conditions and periodontal status in older individuals. *Spec Care Dentist, 38*(6), 373-381. doi:10.1111/scd.12319

21. Chatzopoulos, G. S., Koidou, V. P., Lunos, S., & Wolff, L. F. (2018). Implant and root canal treatment: Survival rates and factors associated with treatment outcome. *J Dent, 71*, 61-66. doi:10.1016/j.jdent.2018.02.005

22. Collares, K., Opdam, N. J. M., Laske, M., Bronkhorst, E. M., Demarco, F. F., Correa, M. B., & Huysmans, M. (2017). Longevity of Anterior Composite Restorations in a General Dental Practice-Based Network. *J Dent Res, 96*(10), 1092-1099. doi:10.1177/0022034517717681

23. Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., .Singleton, P. (2013). Electronic health records: new opportunities for clinical research. *Journal of internal medicine, 274*, 547-560.

24. Corraini, P., Baelum, V., & Lopez, R. (2013). Reliability of direct and indirect clinical attachment level measurements. *J Clin Periodontol, 40*(9), 896-905. doi:10.1111/jcpe.12137

25. Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., .Leenay, M. (2017). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology, 106*, 1-9.

26. Doods, J., Botteri, F., Dugas, M., & Fritz, F. (2014). A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials, 15*, 18. doi:10.1186/1745-6215-15-18

27. Doods, J., Lafitte, C., Ulliac-Sagnes, N., Proeve, J., Botteri, F., Walls, R., .Fritz, F. (2015). A European inventory of data elements for patient recruitment. *Stud Health Technol Inform, 210*, 506-510.

28. Doods, J., Neuhaus, P., & Dugas, M. (2016). Converting ODM Metadata to FHIR Questionnaire Resources. *Stud Health Technol Inform, 228*, 456-460.

29. Dye, B. A., Tan, S., Smith, V., Lewis, B. G., Barker, L. K., Thornton-Evans, G., .Li, C. H. (2007). Trends in oral health status: United States, 1988-1994 and 1999-2004. *Vital Health Stat 11*(248), 1-92.

30. Eke, P. I., Dye, B. A., Wei, L., Slade, G. D., Thornton-Evans, G. O., Borgnakke, W. S., .Genco, R. J. (2015). Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J Periodontol, 86*, 611-622.

31. Eke, P. I., Dye, B. A., Wei, L., Thornton-Evans, G. O., & Genco, R. J. (2012). Prevalence of periodontitis in adults in the United States: 2009 and 2010. *J Dent Res, 91*, 914-920. doi:10.1177/0022034512457373

32. Eke, P. I., Page, R. C., Wei, L., Thornton-Evans, G., & Genco, R. J. (2012). Update of the case definitions for population-based surveillance of periodontitis. *J Periodontol, 83*(12), 1449-1454. doi:10.1902/jop.2012.110664

33. Eke, P. I., Thornton-Evans, G. O., Wei, L., Borgnakke, W. S., Dye, B. A., & Genco, R. J. (2018). Periodontitis in US Adults: National Health and Nutrition Examination Survey 2009-2014. *J Am Dent Assoc, 149*(7), 576-588.e576. doi:10.1016/j.adaj.2018.04.023

34. Eke, P. I., Wei, L., Borgnakke, W. S., Thornton-Evans, G., Zhang, X., Lu, H., .Genco, R. J. (2016). Periodontitis prevalence in adults≥ 65 years of age, in the USA. *Periodontology 2000, 72*, 76-95.

35. Eke, P. I., Wei, L., Thornton-Evans, G. O., Borrell, L. N., Borgnakke, W. S., Dye, B., & Genco, R. J. (2016). Risk indicators for periodontitis in US adults: NHANES 2009 to 2012. *J Periodontol*.

36. Feder, S. L. (2018). Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res, 40*(5), 753-766. doi:10.1177/0193945916689084

37. Foley, T., & Fairmichael, F. (2015). The Potential of Learning Healthcare Systems. *Newcastle: Learning Healthcare Project*.

38. Fransson, H., Dawson, V. S., Frisk, F., Bjorndal, L., & Kvist, T. (2016). Survival of Root-filled Teeth in the Swedish Adult Population. *J Endod, 42*(2), 216-220. doi:10.1016/j.joen.2015.11.008

39. Genco, R. J. (1996). Current view of risk factors for periodontal diseases. *J Periodontol, 67*(10 Suppl), 1041-1049. doi:10.1902/jop.1996.67.10.1041

40. Genco, R. J., & Borgnakke, W. S. (2013). Risk factors for periodontal disease. *Periodontology 2000, 62*, 59-94.

41. Genco, R. J., & Genco, F. D. (2014). Common risk factors in the management of periodontal and associated systemic diseases: the dental setting and interprofessional collaboration. *Journal of Evidence Based Dental Practice, 14*, 4-16.

42. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association, 24*, 198-208.

43. Haber, J., Hartnett, E., Allen, K., Crowe, R., Adams, J., Bella, A., .Vasilyeva, A. (2017). The Impact of Oral-Systemic Health on Advancing Interprofessional Education Outcomes. *Journal of Dental Education, 81*, 140-148.

44. Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). Development of non-invasive diabetes risk prediction models as decision support tools designed for application in the dental clinical environment. *Informatics in Medicine Unlocked, 17*. doi:10.1016/j.imu.2019.100254

45. Hersh, W. R., Cimino, J., Payne, P. R. O., Embi, P., Logan, J., Weiner, M., .Hartzog, T. (2013). Recommendations for the use of operational electronic health record data in comparative effectiveness research. *EGEMS, 1*.

46. Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., .Cimino, J. J. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical care, 51*, S30.

47. Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., & Melton, G. B. (2011). Bias associated with mining electronic health records. *Journal of biomedical discovery and collaboration, 6*, 48.

48. Ingram, W. M., Baker, A. M., Bauer, C. R., Brown, J. P., Goes, F. S., Larson, S., & Zandi, P. P. (2020). Defining Major Depressive Disorder Cohorts Using the EHR: Multiple Phenotypes Based on ICD-9 Codes and Medication Orders. *Neurol Psychiatry Brain Res, 36*, 18-26. doi:10.1016/j.npbr.2020.02.002

49. Jansson, H., & Norderyd, O. (2001). Evaluation of a periodontal risk assessment model in subjects with severe periodontitis. *Swedish dental journal, 25*, 1.

50. Jansson, H., Wahlin, Å., Johansson, V., Åkerman, S., Lundegren, N., Isberg, P.-E., & Norderyd, O. (2014). Impact of Periodontal Disease Experience on Oral Health–Related Quality of Life. *J Periodontol, 85*, 438-445.

51. Consensus training: An effective tool to minimize variations in periodontal diagnosis and treatment planning among dental faculty and students, 77 1022-1032 (2013).

52. Köpcke, F., Trinczek, B., Majeed, R. W., Schreiweis, B., Wenk, J., Leusch, T., .Prokosch, H.-U. (2013). Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Medical Informatics and Decision Making, 13*, 37. doi:10.1186/1472-6947-13-37

53. Kotsakis, G. A., Lian, Q., Ioannou, A. L., Michalowicz, B. S., John, M. T., & Chu, H. (2018). A network meta-analysis of interproximal oral hygiene methods in the reduction of clinical indices of inflammation. *J Periodontol, 89*(5), 558-570. doi:10.1002/jper.17-0368

54. Kumar, S. V., Bangar, S., Neumann, A., Kookal, K. K., Yansane, A., Tokede, O., .Walji, M. (2018). Assessing the validity of existing dental sealant quality measures. *J Am Dent Assoc, 149*(9), 756-764.e751. doi:10.1016/j.adaj.2018.05.001

55. Lane, B. A., Luepke, P., Chaves, E., Maupome, G., Eckert, G. J., Blanchard, S., & John, V. (2015). Assessment of the calibration of periodontal diagnosis and treatment planning among dental students at three dental schools. *J Dent Educ, 79*(1), 16-24.

56. Lang, N. P., Joss, A., Orsanic, T., Gusberti, F. A., & Siegrist, B. E. (1986). Bleeding on probing. A predictor for the progression of periodontal disease? *J Clin Periodontol, 13*(6), 590-596. doi:10.1111/j.1600-051x.1986.tb00852.x

57. Lang, N. P., Suvan, J. E., & Tonetti, M. S. (2015). Risk factor assessment tools for the prevention of periodontitis progression a systematic review. *Journal of clinical periodontology, 42*.

58. Lang, N. P., & Tonetti, M. S. (2003). Periodontal risk assessment (PRA) for patients in supportive periodontal therapy (SPT). *Oral Health Prev Dent, 1*, 7-16.

59. Lee, A. H., Cheung, G. S., & Wong, M. C. (2012). Long-term outcome of primary non-surgical root canal treatment. *Clin Oral Investig, 16*(6), 1607-1617. doi:10.1007/s00784-011-0664-2

60. Levitin, S. A., Grbic, J. T., & Finkelstein, J. (2019). Completeness of Electronic Dental Records in a Student Clinic: Retrospective Analysis. *JMIR Med Inform, 7*(1), e13008. doi:10.2196/13008

61. Lindhe, J., Haffajee, A. D., & Socransky, S. S. (1983). Progression of periodontal disease in adult subjects in the absence of periodontal therapy. *J Clin Periodontol, 10*(4), 433-442. doi:10.1111/j.1600-051x.1983.tb01292.x

62. Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., .Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings, 2013*, 149.

63. Lobach, D. F., & Detmer, D. E. (2007). Research challenges for electronic health records. *American Journal of Preventive Medicine, 32*(5), S104-S111.

64. Loe, H., Anerud, A., Boysen, H., & Morrison, E. (1986). Natural history of periodontal disease in man. Rapid, moderate and no loss of attachment in Sri Lankan laborers 14 to 46 years of age. *J Clin Periodontol, 13*(5), 431-445. doi:10.1111/j.1600-051x.1986.tb01487.x

65. Machtei, E. E., Hausmann, E., Dunford, R., Grossi, S., Ho, A., Davis, G., .Genco, R. J. (1999). Longitudinal study of predictive factors for periodontal disease and tooth loss. *J Clin Periodontol, 26*(6), 374-380. doi:10.1034/j.1600-051x.1999.260607.x

66. Marlow, A. K., Hamada, Y., Maupome, G., Eckert, G. J., & John, V. (2018). Periodontal diagnosis and treatment planning among Indiana dental faculty, periodontists, and general practice dentists: A multi-group comparison. *Journal of Dental Education, 82*, 291-298. doi:10.21815/JDE.018.029

67. Martin, S., Wagner, J., Lupulescu-Mann, N., Ramsey, K., Cohen, A., Graven, P., .Dorr, D. A. (2017). Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. *Appl Clin Inform, 8*(3), 794-809. doi:10.4338/aci-2016-12-ra-0210

68. Matuliene, G., Studer, R., Lang, N. P., Schmidlin, K., Pjetursson, B. E., Salvi, G. E., .Zwahlen, M. (2010). Significance of periodontal risk assessment in the recurrence of periodontitis and tooth loss. *Journal of clinical periodontology, 37*, 191-199.

69. Maupome, G., & Sheiham, A. (2002). Explanatory models in the interpretations of clinical features of dental patients within a university dental education setting. *European Journal of Dental Education, 6*, 2-8. doi:10.1034/j.1600-0579.2002.060102.x

70. McFall, W. T., Jr. (1982). Tooth loss in 100 treated patients with periodontal disease. A long-term study. *J Periodontol, 53*(9), 539-549. doi:10.1902/jop.1982.53.9.539

71. Mertz, E., Bolarinwa, O., Wides, C., Gregorich, S., Simmons, K., Vaderhobli, R., & White, J. (2015). Provider attitudes toward the implementation of clinical decision support tools in dental practice. *Journal of Evidence Based Dental Practice, 15*, 152-163.

72. Mertz, E., Wides, C., & White, J. (2017). Clinician attitudes, skills, motivations and experience following the implementation of clinical decision support tools in a

large dental practice. *J Evid Based Dent Pract, 17*(1), 1-12. doi:10.1016/j.jebdp.2016.10.001

73. Meyer-Baumer, A., Pritsch, M., Cosgarea, R., El Sayed, N., Kim, T. S., Eickholz, P., & Pretzl, B. (2012). Prognostic value of the periodontal risk assessment in patients with aggressive periodontitis. *J Clin Periodontol, 39*(7), 651-658. doi:10.1111/j.1600-051X.2012.01895.x

74. Montagner, A. F., Sande, F. H. V., Muller, C., Cenci, M. S., & Susin, A. H. (2018). Survival, Reasons for Failure and Clinical Characteristics of Anterior/Posterior Composites: 8-Year Findings. *Braz Dent J, 29*(6), 547-554. doi:10.1590/0103-6440201802192

75. Morelli, T., Moss, K. L., Preisser, J. S., Beck, J. D., Divaris, K., Wu, D., & Offenbacher, S. (2018). Periodontal profile classes predict periodontal disease progression and tooth loss. *J Periodontol, 89*(2), 148-156. doi:10.1002/jper.17-0427

76. Morley, K. I., Wallace, J., Denaxas, S. C., Hunter, R. J., Patel, R. S., Perel, P., .Hemingway, H. (2014). Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One, 9*(11), e110900. doi:10.1371/journal.pone.0110900

77. Mullins, J. M., Even, J. B., & White, J. M. (2016). Periodontal Management by Risk Assessment: A Pragmatic Approach. *J Evid Based Dent Pract, 16 Suppl*, 91-98. doi:10.1016/j.jebdp.2016.01.020

78. Mullins, J. M., Even, J. B., & White, J. M. (2016). Periodontal management by risk assessment: a pragmatic approach. *Journal of Evidence Based Dental Practice, 16*, 91-98.

79. Murray, C. J. L., Barber, R. M., Foreman, K. J., Ozgoren, A. A., Abd-Allah, F., Abera, S. F., .Abu-Raddad, L. J. (2015). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition. *The Lancet, 386*, 2145-2191.

80. Muthee, V., Bochner, A. F., Osterman, A., Liku, N., Akhwale, W., Kwach, J., .Puttkammer, N. (2018). The impact of routine data quality assessments on electronic medical record data quality in Kenya. *PLoS One, 13*(4), e0195362. doi:10.1371/journal.pone.0195362

81. Needleman, I., Nibali, L., & Di Iorio, A. (2015). Professional mechanical plaque removal for prevention of periodontal diseases in adults--systematic review update. *J Clin Periodontol, 42 Suppl 16*, S12-35. doi:10.1111/jcpe.12341

82. Nibali, L., Sun, C., Akcalı, A., Meng, X., Tu, Y. K., & Donos, N. (2017). A retrospective study on periodontal disease progression in private practice. *Journal of Clinical Periodontology, 44*, 290-297. doi:10.1111/jcpe.12653

83. Noorden, H. L. R. V. (05 JUNE 2020). High-profile coronavirus retractions raise concerns about data oversight. *Nature*. Retrieved from https://www.nature.com/articles/d41586-020-01695-w

84. Ogawa, H., Yoshihara, A., Hirotomi, T., Ando, Y., & Miyazaki, H. (2002). Risk factors for periodontal disease progression among elderly people. *J Clin Periodontol, 29*(7), 592-597. doi:10.1034/j.1600-051x.2002.290702.x

85. Oliveira Costa, F., Cota, L. O., Costa, J. E., & Pordeus, I. A. (2007). Periodontal disease progression among young subjects with no preventive dental care: a 52-month follow-up study. *J Periodontol, 78*(2), 198-203. doi:10.1902/jop.2007.060150

86. Page, R. C., & Eke, P. I. (2007). Case definitions for use in population-based surveillance of periodontitis. *J Periodontol, 78*(7 Suppl), 1387-1399. doi:10.1902/jop.2007.060264

87. Page, R. C., Martin, J., Krall, E. A., Mancl, L., & Garcia, R. (2003). Longitudinal validation of a risk calculator for periodontal disease. *Journal of clinical periodontology, 30*, 819-827.

88. Page, R. C., Martin, J. A., & Loeb, C. F. (2004). Use of risk assessment in attaining and maintaining oral health. *Compendium of continuing education in dentistry (Jamesburg, NJ: 1995), 25*, 657-660, 663-656, 669; quiz 670.

89. Page, R. C., Martin, J. A., & Loeb, C. F. (2005). The Oral Health Information Suite (OHIS): its use in the management of periodontal disease. *Journal of Dental Education, 69*, 509-520.

90. Papapanou, P. N., & Wennstrom, J. L. (1990). A 10-year retrospective study of periodontal disease progression. Clinical characteristics of subjects with pronounced and minimal disease development. *J Clin Periodontol, 17*(2), 78-84. doi:10.1111/j.1600-051x.1990.tb01066.x

91. Patel, J., Mowery, D., Krishnan, A., & Thyvalikakath, T. (2018). Assessing Information Congruence of Documented Cardiovascular Disease between Electronic Dental and Medical Records. *AMIA Annu Symp Proc, 2018*, 1442-1450.

92. Patel, J., Siddiqui, Z., Krishnan, A., & Thyvalikakath, T. (2017). Identifying Patients' Smoking Status from Electronic Dental Records Data. *Stud Health Technol Inform, 245*, 1281.

93. Patel, J., Siddiqui, Z., Krishnan, A., & Thyvalikakath, T. P. (2018). Leveraging Electronic Dental Record Data to Classify Patients Based on Their Smoking Intensity. *Methods Inf Med, 57*(5-06), 253-260. doi:10.1055/s-0039-1681088

94. Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc, 20*(e2), e206-211. doi:10.1136/amiajnl-2013-002428

95. Perez-Chaparro, P. J., Goncalves, C., Figueiredo, L. C., Faveri, M., Lobao, E., Tamashiro, N., .Feres, M. (2014). Newly identified pathogens associated with periodontitis: a systematic review. *J Dent Res, 93*(9), 846-858. doi:10.1177/0022034514542468

96. Persson, G. R., Mancl, L. A., Martin, J., & Page, R. C. (2003). Assessing periodontal disease risk: a comparison of clinicians' assessment versus a computerized tool. *The Journal of the American Dental Association, 134*, 575-582.

97. Pirani, C., Chersoni, S., Montebugnoli, L., & Prati, C. (2015). Long-term outcome of non-surgical root canal treatment: a retrospective analysis. *Odontology, 103*(2), 185-193. doi:10.1007/s10266-014-0159-0

98. Raedel, M., Priess, H. W., Bohm, S., Noack, B., Wagner, Y., & Walter, M. H. (2019). Tooth loss after periodontal treatment—Mining an insurance database. *Journal of Dentistry, 80*, 30-35. doi:10.1016/j.jdent.2018.11.001

99. Ramseier, C. A., Anerud, A., Dulac, M., Lulic, M., Cullinan, M. P., Seymour, G. J., .Lang, N. P. (2017). Natural history of periodontitis: Disease progression and tooth loss over 40 years. *J Clin Periodontol, 44*(12), 1182-1191. doi:10.1111/jcpe.12782

100. Reddy, M. S., Geurs, N. C., Jeffcoat, R. L., Proskin, H., & Jeffcoat, M. K. (2000). Periodontal disease progression. *J Periodontol, 71*(10), 1583-1590. doi:10.1902/jop.2000.71.10.1583

101. Reimer, A. P., Milinovich, A., & Madigan, E. A. (2016). Data quality assessment framework to assess electronic medical record data for use in research. *Int J Med Inform, 90*, 40-47. doi:10.1016/j.ijmedinf.2016.03.006

102. Sai Sujai, G. V., Triveni, V. S., Barath, S., & Harikishan, G. (2015). Periodontal risk calculator versus periodontal risk assessment. *J Pharm Bioallied Sci, 7*(Suppl 2), S656-659. doi:10.4103/0975-7406.163593

103. Saver, J. L., Warach, S., Janis, S., Odenkirchen, J., Becker, K., Benavente, O., .Schwamm, L. (2012). Standardizing the structure of stroke clinical and epidemiologic research data: the National Institute of Neurological Disorders and Stroke (NINDS) Stroke Common Data Element (CDE) project. *Stroke, 43*(4), 967-973. doi:10.1161/strokeaha.111.634352

104. Schatzle, M., Faddy, M. J., Cullinan, M. P., Seymour, G. J., Lang, N. P., Burgin, W., .Loe, H. (2009). The clinical course of chronic periodontitis: V. Predictive factors in periodontal disease. *J Clin Periodontol, 36*(5), 365-371. doi:10.1111/j.1600-051X.2009.01391.x

105. Schatzle, M., Loe, H., Burgin, W., Anerud, A., Boysen, H., & Lang, N. P. (2003). Clinical course of chronic periodontitis. I. Role of gingivitis. *J Clin Periodontol, 30*(10), 887-901. doi:10.1034/j.1600-051x.2003.00414.x

106. Schatzle, M., Loe, H., Lang, N. P., Burgin, W., Anerud, A., & Boysen, H. (2004). The clinical course of chronic periodontitis. *J Clin Periodontol, 31*(12), 1122-1127. doi:10.1111/j.1600-051X.2004.00634.x

107. Schatzle, M., Loe, H., Lang, N. P., Heitz-Mayfield, L. J., Burgin, W., Anerud, A., & Boysen, H. (2003). Clinical course of chronic periodontitis. III. Patterns, variations and risks of attachment loss. *J Clin Periodontol, 30*(10), 909-918. doi:10.1034/j.1600-051x.2003.00401.x

108. Schatzle, M., Loe, H., Ramseier, C. A., Burgin, W., Anerud, A., Boysen, H., & Lang, N. P. (2010). Clinical course of chronic periodontitis: effect of lifelong light smoking (20 years) on loss of attachment and teeth. *J Investig Clin Dent, 1*(1), 8-15. doi:10.1111/j.2041-1626.2010.00008.x

109. Schlegel, D. R., & Ficheur, G. (2017). Secondary Use of Patient Data: Review of the Literature Published in 2016. *Yearb Med Inform, 26*(1), 68-71. doi:10.15265/iy-2017-032

110. Schleyer, T., Song, M., Gilbert, G. H., Rindal, D. B., Fellows, J. L., Gordan, V. V., & Funkhouser, E. (2013). Electronic dental record use and clinical information management patterns among practitioner-investigators in The Dental Practice-

Based Research Network. *J Am Dent Assoc, 144*(1), 49-58. doi:10.14219/jada.archive.2013.0013

111. Shen, F., Liu, S., Wang, Y., Wen, A., Wang, L., & Liu, H. (2018). Utilization of Electronic Medical Records and Biomedical Literature to Support the Diagnosis of Rare Diseases Using Data Fusion and Collaborative Filtering Approaches. *JMIR Med Inform, 6*(4), e11301. doi:10.2196/11301

112. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc, 21*(2), 221-230. doi:10.1136/amiajnl-2013-001935

113. Singer, A., Yakubovich, S., Kroeker, A. L., Dufault, B., Duarte, R., & Katz, A. (2016). Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? *J Am Med Inform Assoc, 23*(6), 1107-1112. doi:10.1093/jamia/ocw013

114. Skyttberg, N., Chen, R., Blomqvist, H., & Koch, S. (2017). Exploring Vital Sign Data Quality in Electronic Health Records with Focus on Emergency Care Warning Scores. *Appl Clin Inform, 8*(3), 880-892. doi:10.4338/aci-2017-05-ra-0075

115. Song, M., Liu, K., Abromitis, R., & Schleyer, T. L. (2013). Reusing electronic patient data for dental clinical research: a review of current status. *J Dent, 41*(12), 1148-1163. doi:10.1016/j.jdent.2013.04.006

116. Sperrin, M., Thew, S., Weatherall, J., Dixon, W., & Buchan, I. (2011). Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. *AMIA Annu Symp Proc, 2011*, 1318-1325.

117. Spratt, S. E., Pereira, K., Granger, B. B., Batch, B. C., Phelan, M., Pencina, M., .Jelesoff, N. (2017). Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J Am Med Inform Assoc, 24*(e1), e121-e128. doi:10.1093/jamia/ocw123

118. Taruscio, D., Mollo, E., Gainotti, S., Posada de la Paz, M., Bianchi, F., & Vittozzi, L. (2014). The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration. *Arch Public Health, 72*(1), 35. doi:10.1186/2049-3258-72-35

119. Thyvalikakath, T., Song, M., & Schleyer, T. (2018). Perceptions and attitudes toward performing risk assessment for periodontal disease: a focus group exploration. *BMC Oral Health, 18*(1), 90. doi:10.1186/s12903-018-0550-2

120. Thyvalikakath, T. P., Duncan, W. D., Siddiqui, Z., LaPradd, M., Eckert, G., Schleyer, T., .Gilbert, G. H. (2020). Leveraging Electronic Dental Record Data for Clinical Research in the National Dental PBRN Practices. *Appl Clin Inform, 11*(2), 305-314. doi:10.1055/s-0040-1709506

121. Thyvalikakath TP, D. W., Siddiqui Z, LaPradd M, Jurkovich M, Shea TL, Bogacz D, Yu T, Fellow JL, Gordan VV, Gilbert GH, National Dental PBRN Collaborator Group et. al. (2019). Survival analysis of endodontically treated teeth in National Dental Practice-Based Network practices. *J Dent Res 2019 98, (A)*(0510).

122. Thyvalikakath TP, D. W., Siddiqui Z, LaPradd M, Jurkovich M, Shea TL, Bogacz D, Yu T, Fellow JL, Gordan VV, Gilbert GH, National Dental PBRN Collaborator Group et. al. (2020). Survival Analysis of Posterior Composite Restorations in National-Dental-PBRN Practices. *American Association of Dental Research General Session & Exhibition*.

123. Tonetti, M. S., Eickholz, P., Loos, B. G., Papapanou, P., Velden, U., Armitage, G., .Hughes, F. (2015). Principles in prevention of periodontal diseases. *Journal of clinical periodontology, 42*.

124. Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J Periodontol, 89 Suppl 1*, S159-s172. doi:10.1002/jper.18-0006

125. Trombelli, L., Farina, R., Ferrari, S., Pasetti, P., & Calura, G. (2009). Comparison between two methods for periodontal risk assessment. *Minerva Stomatol, 58*, 277-287.

126. Trombelli, L., Farina, R., Silva, C. O., & Tatakis, D. N. (2018). Plaque-induced gingivitis: Case definition and diagnostic considerations. *J Periodontol, 89 Suppl 1*, S46-s73. doi:10.1002/jper.17-0576

127. van der Lei, J. (1991). Use and abuse of computer-stored medical records. *Methods Inf Med, 30*(2), 79-80.

128. Van Dyke, T. E., & Dave, S. (2005). Risk factors for periodontitis. *Journal of the International Academy of Periodontology, 7*, 3.

129. Violan, C., Foguet-Boreu, Q., Hermosilla-Perez, E., Valderas, J. M., Bolibar, B., Fabregas-Escurriola, M., .Munoz-Perez, M. A. (2013). Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. *BMC Public Health, 13*, 251. doi:10.1186/1471-2458-13-251

130. Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., .Coates, M. M. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*, 1459-1544.

131. Wang, S. Y., Wang, Z. D., & Yan, B. (2018). [Advances in surgical techniques of periodontal corticotomy]. *Hua Xi Kou Qiang Yi Xue Za Zhi, 36*(2), 220-225. doi:10.7518/hxkq.2018.02.020

132. Wang, Y., Siddiqui, Z., Krishnan, A., Patel, J., & Thyvalikakath, T. (2017). Extraction and Evaluation of Medication Data from Electronic Dental Records. *Stud Health Technol Inform, 245*, 1290.

133. Weiskopf, N. G., Bakken, S., Hripcsak, G., & Weng, C. (2017). A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC), 5*(1), 14. doi:10.5334/egems.218

134. Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc, 20*(1), 144-151. doi:10.1136/amiajnl-2011-000681

135. Wiebe, C. B., & Putnins, E. E. (2000). The periodontal disease classification system of the American Academy of Periodontology--an update. *J Can Dent Assoc, 66*(11), 594-597.

136. The periodontal disease classification system of the American Academy of Periodontology - An update, 66 594-597 (2000).

137. Wu, M., Chen, S.-W., & Jiang, S.-Y. (2015). Relationship between Gingival Inflammation and Pregnancy. *Mediators of Inflammation, 2015*, 11. doi:10.1155/2015/623427

Curriculum Vitae

Jay Sureshbhai Patel

Education

Indiana University, IN: PhD, Health Informatics	2015-2020
Rutgers University, NJ: MS, Biomedical Informatics	2014-2015
Rajiv Gandhi University, India: Bachelor of Dental Surgery	2006-2013

Published Papers

1. Patel, J., Siddiqui, Z., Krishnan, A., & Thyvalikakath, T. P. (2018). Leveraging Electronic Dental Record Data to Classify Patients Based on Their Smoking Intensity. Methods of information in medicine.

2. Patel J, Mowery D, Krishnan A, Thyvalikakath T. (2018). Assessing information congruence of documented cardiovascular disease between electronic dental and medical records. American medical informatics association conference proceedings.

3. Holden, R. J., Binkheder, S., Patel, J., & Viernes, S. H. P. (2018). Best Practices for Health Informatician Involvement in Interprofessional Health Care Teams. Applied clinical informatics, 9(01), 141-148.

Abstracts

1. Patel J, Zai H, Thyvalikakath TP. Utilizing Electronic Dental Record Data to Monitor Periodontal Disease Progression. MCBK, abstract 2019.
2. Medam J, Williams K, Patel J, Thyvalikakath TP. Reasons, Information, and Time: Exploration of Dental Clinician-Initiated Medical Consultations, AMIA 2019 Annual Symposium.

3. Medam J, Williams K, Patel J, Gonzales T, Thyvalikakath T. Qualitative Exploration of Factors Associated with Dental Provider Initiated Medical Consultations. J Dent Res Vol C: 2018. 0002. (www.iadr.org)
4. Patel J, Siddiqui Z, Krishnan A, Thyvalikakath T. Identifying patients' smoking status from electronic dental records data. Stud Health Technol Inform. 2017; 245:1281.
5. Wang Y, Siddiqui, Krishnan A, Patel J, Thyvalikakath T. Extraction and evaluation of medication data from Electronic Dental Records. Stud Health Technol Inform. 2017; 245:1290.
6. Patel J, Krishnan A, Mowery D, Thyvalikakath T. Automated Identification of Patient-reported Cardiovascular Diseases from Electronic Dental Records J Dent Res 96 (Spec Iss A): 3070, 2017. (www.iadr.org)
7. Wang Y, Siddiqui Z, Patel J, Thyvalikakath T. Medication Profile of Dental Patients in an Academic Setting. J Dent Res 96 (Spec Iss A): 2514, 2017. (www.iadr.org)
8. Orenstein D, Dhankhar U, Patel J, Gonzalez T, Oldham J, Krishnan A, Thyvalikakath T. Developing a Gold Standard to Identify Reasons for Denture Remakes. J Dent Res 96 (Spec Iss A): 0171, 2017. (www.iadr.org)
9. Patel J, Siddiqui Z, Mowery D, Thyvalikakath TP. Annotating patient's smoking status from electronic dental record histories. AMIA 2016 Annual Symposium, Chicago, IL; November 12-16, 2016.
10. Jones J, Wu H, Patel J, Kasthurirathne S, An Evaluation of Activity Tracker for Monitoring Parkinson's Disease Patient Outcomes. AMIA 2016 Annual Symposium, Chicago, IL; November 12-16, 2016.

11. Patel J, Mowery D, Thyvalikakath TP. Representing and annotating coronary artery disease from patient's medical history. J Dent Res 95 Spec Iss: B, abstract 1013, 2016. (www.iadr.org)
12. Radler D, Marcus A, Patel J, Griegns R, Changes in Dietary Intake among Participants in a University Worksite Wellness Program Using the Dietary Screening Questionnaire, 2015. Rutgers Sch. Of Hlth. Related Professions.

Professional Memberships

American Medical Informatics Association: 2016-present

Medical Informatics Conference: 2016,2017

Indy Big Data Conference: 2017, 2018

American Association of Dental Research: 2015, 2016

Mobilizing Computable Biomedical Knowledge: 2018