Application of Data Science to Discover the Relationship between Dental Caries and Diabetes in Dental Records

Germán H. Alférez[1], Jany Jiménez[2], Héctor Hernández Navarro[3], Merari González[4], Rusbel Domínguez[5], Alejandro Briones[6], Héctor Hernández Villalvazo[7]

harveyalferez@um.edu.mx[1], jany.jimenez@um.edu.mx[2], odontos@um.edu.mx[3], iscmera@gmail.com[4], ruxbel@gmail.com[5], alexbriones211191@gmail.com[6], hector.hdz.villalvazo@gmail.com[7]
Universidad de Montemorelos, Apartado 16-5,
Montemorelos N.L. 67500, Mexico

Abstract - *Diabetes is a chronic and metabolic disease. According to the World Health Organization (WHO), 422 million of adults suffer from diabetes worldwide. In fact, in 2012 diabetes caused 1.5 million deaths in the world. In Mexico, our country, diabetes is a highly relevant public health problem. For example, in 2015 there were 11 million cases of diabetes. Our contribution is to apply novel data science techniques to medical records at a dental clinic in Northeast Mexico to discover the relationship between diabetes and dental caries. Our research work follows IBM's data science methodology. The analysis of data was carried out with machine learning. Five experiments were performed on 193 dental records. Our findings corroborate the results in related work.*

Keywords: Diabetes, Dental Caries, Data Science

1) Introduction

Diabetes is a chronic and metabolic disease. It is characterized by high blood sugar levels that result from the deficit of the body to produce insulin. Insulin is the hormone that regulates the body glucose usage. Diabetes leads to damages in the heart, blood vessels, eyes, kidneys, and nerves.

According to the World Health Organization (WHO), 422 million of adults suffer from diabetes worldwide. In Mexico, our country, diabetes has been the first cause of death in women and the second in men since the year 2000.

Previous research work has shown that diabetes is closely related to dental caries since both of them share similar risk factors. For instance, diabetic patients that do not have control of their blood sugar levels have a higher risk of presenting systemic and oral complications [1]. One of these complications is dental caries. Dental caries is a multifactorial progressive process that can be developed on the tooth surface, inside the oral cavity, where the plaque allows it to grow on a period of time.

Diabetic patients are two or three times more susceptible to develop a periodontal disease than healthy patients [2]. Moreover, diabetes is the first cause of premature tooth loss, interrupting the physiological function of mastication, leading to a softer diet with a higher level of sugar that can cause dysglycemia [3].

Our contribution is to apply novel data science techniques to medical records to discover the relationship between diabetes and dental caries. Data science consists of analyzing data, structured and unstructured, to get knowledge [4]. Although the relationship between diabetes and dental caries has been found in related work, previous approaches are based on field studies with a limited number of patients. This may be caused by the lack of automatized techniques to analyze data in dentistry. Our approach goes a step further with the analysis of a larger number of patients' data by means of machine learning techniques.

Specifically, this work focuses on finding hidden patterns in 193 dental records of patients at the Dental Clinic "Luz y Vida" located at the campus of Universidad de Montemorelos, Montemorelos, N.L., Mexico. This study follows IBM's data science methodology [5]. In the experiments, the k-means algorithm of machine learning was executed in Weka. Seven features in the clinical records were analyzed.

This paper is organized as follows. The second section presents the description of the dataset taken from de clinical records at Dental Clinic "Luz y Vida". The third section presents how we have

applied IBM's methodology to find hidden patterns in the dataset to discover the relationship between dental caries and diabetes. The last section presents conclusions and future work.

2) Description of the Dataset

In this research work, the data was obtained from 193 clinical records from the Dental Clinic "Luz y Vida". This clinic is located at the campus of Universidad de Montemorelos, Montemorelos, N.L., Mexico. Since its foundation in 2004, this clinic has stored dental records in paper forms.

The clinical records used in this experiment belong to 193 patients located at the following cities: Montemorelos (140 patients), Monterrey (2 patients), Linares (4 patients), Solistahuacan (1 patient), Allende (9 patients), Mexico City (1 patient), Alamo (1 patient), Tampico (1 patient), Hermosillo (1 patient), Tamaulipas (1 patient), Coatzacoalcos (1 patient), Cancun (1 patient), Cd. Madero (1 patient), Tamasopo (1 patient), Altamira (1 patient), General Terán (3 patients), Anahuac (1 patient), Chihuahua (1 patient), Elkhart (1 patient), Santiago Tuxtla (1 patient), Camargo (1 patient), Cadereyta (1 patient), Chula Vista (1 patient), Mezcalapa (1 patient), Rayones (1 patient), Navojoa (1 patient), Georgia (1 patient), Riverside (1 patient), Caborca (1 patient). The location of 11 patients was not recorded in the dental records. Since this clinic has changed several times the paper forms used to record clinical records, we decided to take the sample for the experiments from the latest 193 clinical records, which use the same paper form.

3) Applying IBM's Methodology to Clinical Records

Data science requires a methodology that eases its application to industry and academy. That is why IBM offers a methodology for the application of data science [5]. This methodology is organized in ten stages structured in an iterative process. In the following paragraphs we describe the stages that were followed in this research according to IBM's data science methodology:

1. Understanding the problem: Dental caries related to diabetes is a hot topic in dentistry because there is not a clear and absolute position about the relationship between these two diseases. Moreover, related work tends to expose this relationship through field studies with a limited number of patients. In these studies, researchers take a sample of patients with diabetes and then look for the possible relationship between diabetes and dental caries. However, these studies do not propose the automatized analysis of dental records to facilitate the process.

For example, Seethalakshmi et al. [6] evaluated oral diseases that can be caused by diabetes, such as the incidence of dental caries and salivary pH in 20 patients. The results showed that not only periodontal health was affected, but also salivary pH had a decrease of 6.51. Likewise, the incidence of dental caries increased significantly in comparison with the patients without diabetes.

On one hand, Novotna et al. [7] mention that there is an increase of plaque levels and chronic gingivitis as much in adults as in children with type 1 diabetes. On the other hand, Miranda et al. [8] found out that oral health in patients with type 1 diabetes in Chile is more precarious than in healthy patients. However, they mention that this problem could be caused by poor hygiene.

Singh et al. [9] came to the conclusion that patients with type 2 diabetes have a higher risk of developing dental caries. Also, they pointed out that saliva flow and saliva calcium levels are significantly lower compared to healthy patients. Therefore a reduction in the saliva components reduces the enamels capacity to endure the remineralization and demineralization process. It creates the right environment for dental caries.

Oral bacteria are with no doubt a determinant factor in the formation of dental caries. For example, Kampoo et al. [10] found that the incidence in diabetic patients in Thailand is much higher compared to non-diabetics. Also, the number of acidogenic bacteria in diabetic patients is much higher than in healthy patients. Therefore, the high dental caries incidence in diabetic patients in Thailand is positively related to de Streptococcus and Lactobacillus bacteria.

Iqbal et al. [11] made a study to establish if there is a relationship between diabetes mellitus and dental caries by measuring glucose levels related to dental caries in different patients. These authors found out that glucose levels in diabetic patients' saliva are slightly higher than in healthy patients. Also, the levels of calcium in diabetic patients' saliva are lower.

Jawed et al. [12] found out that the level of blood sugar and glycosylated hemoglobin, and the number of decayed, missing, and filled teeth (DMF) is significantly higher in type 2 diabetics than in healthy patients. These results were obtained by a saliva sample and a DMF test (DMFT).

Similarly, Miko et al. [13] mention that the deficiency in glycemic control as well as the early

occurrence of diabetes can increase the risk of dental caries. This study was made with a DMFT applied to 259 teenagers with type 1 diabetes.

Stojanoviç et al. [14] studied the condition of type 2 diabetic patients related to metabolic control. The sample was composed by 47 type 2 diabetic patients randomly chosen and divided into two groups: poorly controlled diabetics and controlled patients. They found out that patients with a poor control of diabetes have a significant higher amount of dental caries compared to those that control the disease.

Hintao et al. [15] found that patients with type 2 diabetes, compared to healthy patients, have a higher risk of root surface dental caries. However, the prevalence and crown surface decay were not significantly different. Therefore, they concluded that type 2 diabetes is an important risk factor for root decay, but not for crown surface decay.

2. Analytic approach: In this stage, machine learning was used to analyze the data from the clinical dental records. Machine learning is a branch of artificial intelligence that consists of developing techniques that allow computers to learn by means of analyzing structured or unstructured data [16].

Weka[1] was used to analyze the dental records by means of machine learning. Weka is a data-mining software developed by the University of Waikato [17]. This software was programmed in Java and has powerful algorithms to extract information contained in datasets [18].

3. Data requirements: The clinical records at the Dental Clinic "Luz y Vida" have 60 features. In the field of machine learning, a feature is a variable that summarizes key aspects to be analyzed. In our case, the features contain data about personal information, anamnesis, and intraoral exploration (see Table 1).

4. Data collection: The clinical records analyzed in this study were in paper. Therefore, 15 students at the School of Engineering and Technology of Universidad de Montemorelos digitalized them. These students took a period of around two months in this process.

5. Data understanding: In this step, we decreased the number of features to 7. These features are the ones that we considered to be associated to dental caries and diabetes according to related work (see the related work presented in the first step). These features are as follows: 1) endocrine problems, including diabetes, family history of diabetes, thyroid gland problems, and others; 2) teeth problems, including sensitivity and bad habits (biting nails, thumb sucking, pencil biting, etc.); 3) number of decayed teeth; 4) number of missing teeth; 5) number of restored or filled teeth; 6) age; and 7) blood type.

6. Data preparation: In this stage, the data related to each feature in the clinical records was converted to numbers. This process was necessary because Weka requires numeric values to do the analysis. The dataset with the studied clinical records is available online[2].

7. Modeling: Among the machine learning algorithms that Weka provides, we chose k-means to generate clusters that model the relationship between dental caries and diabetes in the clinical records. k-means is an unsupervised-learning algorithm that allows to group data in clusters by discovering their centroids [18]. In k-means each sample inside the dataset must be included in one of the clusters [17]. We decided to use this algorithm because of the following reasons: 1) data in the clinical records was not labeled; and 2) k-means is a very popular and effective clustering algorithm [19].

8. Deployment: The experiments were executed with a different number of clusters (k from 4 to 7). We chose the results from the experiments with the minimum within cluster sum of squared errors (sum of distance functions of each point in the cluster to the k center). The results were iteratively presented to the team of dentists to get their feedback and avoid results that were not congruent with their domain of knowledge.

As a way of illustration, Figure 1 shows one of the results in the experiments. First, this figure shows the number of iterations in the experiment (11 in this case) and the within cluster sum of squared errors (21.55317619018957). This image also shows the selected features for each test (21, 30, 45, 54, 56 and 57. Table 1 shows the descriptions of each one of these features), the number of clusters (k = 6), and the number of instances or samples contained in each cluster (e.g. 28 instances in Cluster 0).

The results of the 5 experiments that were conducted on the 193 clinical records are described as follows:

[1] Weka: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[2] Our dataset: https://docs.google.com/spreadsheets/d/1prAv_cj6nFpZejotYje4gpfqjT291vM_tpu9cLTZSGg/edit?usp=sharing

Table 1. Description of dental records

Patient Information	
No.	*description*
1	age
2	gender
3	city
4	state
5	marital status
6	occupation
7	medical service
8	religion
9	health status

Anamnesis			
No.	*description*	*No.*	*description*
10	under medical care	28	other diseases (e.g. asthma, cancer, etc.)
11	hospitalized or sick	29	heart disease
12	excessive bleeding	30	endocrine disease
13	pregnant	31	bones or muscles disease
14	if pregnant, date of birth	32	digestive disease
15	smoking	33	urinary disease
16	alcoholic beverages	34	allergies
17	more than two drinks per day	35	anesthesia reaction
18	drug use	36	problem with dental treatment
19	frequently tired	37	nerves dental treatment
20	skin disease	38	last dental visit
21	eyes disease	39	mouth disease
22	ears disease	40	teeth disease
23	nose disease	41	brushing teeth
24	throat disease	42	toothbrushing
25	nervous system disease	43	flossing
26	respiratory disease	44	fluorine usage
27	blood disease	45	current medicines (e.g. antibiotics, nitroglycerin, aspirin)

Intraoral Scan (Normal or Abnormal)	
No.	*description*
46	lips
47	tongue
48	corners of lips
49	soft palate
50	hard palate
51	mucous membrane
52	floor of the mouth
53	salivary glands
54	number decayed teeth
55	number of teeth with pain
56	number of missing teeth
57	number of restored teeth
58	restoration status
59	periodontal status
60	occlusion status

- **First experiment:** The objective of this experiment was to verify that diabetic patients also present problems in their teeth. 8 clusters were analyzed in this experiment. We found out that diabetic patients tend to present teeth loss and food accumulation in some zones. The within cluster sum of squared errors in this experiment was 2.5.

- **Second experiment:** The objective of this experiment was to analyze the number of teeth with caries in diabetic and healthy patients. We found out that diabetic patients tend to present 9 to 17 teeth with caries, whereas healthy patients tend to present between 1 and 9 teeth with caries. The within cluster sum of squared errors in this experiment was 6.8. The data was organized in 4 clusters.

- **Third experiment:** The objective of this experiment was to analyze the following features: endocrine disease, number of decayed teeth, number of missing teeth, and number of restored teeth. The data was organized in 6 clusters and the within cluster sum of squared errors in this experiment was 14.3. The analysis of these features showed that diabetic patients tend to present between 9 and 17 decayed teeth, a lower number of teeth, and a lower number of restored teeth than healthy patients.

- **Fourth experiment:** The objective of this experiment was to analyze the following features: age, blood type, if the patient presents endocrine problems, and total number of teeth decay. On one hand, we found out that there is a tendency in patients around 23 years old to not present diabetes and to have around 7 teeth decay. On the other hand, patients around 56 years old tend to easily have gum bleeding and tend to be diabetics or have a relative with diabetes. Also, these patients present around 9 decayed teeth. As the age increases, the number of decayed teeth also increases. The within cluster sum of squared errors in this experiment was 11.96. The data was organized in 7 clusters.

- **Fifth experiment:** The objective of this experiment was to analyze the following features: eyes disease (since diabetes can cause vision problems), endocrine disease, if the patient takes medicines, and the number of decayed, missing, and restored teeth. We found out that healthy patients tend to have between 7 and 9 decayed teeth. In this experiment the error range was 21.5. The data was organized in 6 clusters. This finding supports the result obtained in the second experiment: healthy patients have a lower tendency to have dental caries than diabetic patients.

Figure 1. Sample result of an experiment in Weka

kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 21.553176719018957
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(193)	(28)	(52)	(42)	(38)	(18)	(15)
21	0.2487	1	0	0.0714	0.0263	0.6667	0.2667
30	1.1606	0.5357	0.0769	3	1.0263	1.2222	1.2
45	1.8152	0.3797	0.4089	1.881	0.1004	12.7222	0.442
54	9.2721	8.464	7.0242	7.8179	16.7368	8.0502	5.2
56	3.1322	3.7972	3.1648	3.0091	2.1227	3.6217	4.0931
57	3.0909	2.539	2.5087	2.1104	2.012	2.4394	12.4

9. Evaluation: The evaluation of the experiments' results was made by a team of dentists at the Dental Clinic "Luz y Vida". This team analyzed each one of the results obtained in the previous step. Based on the presented results, they concluded that the clinical record analysis of patients' data trough data science corroborates the existence of the relationship between dental caries and diabetes. This results supports the findings of related work in this area. For example, according to Seethalakshmi et al. [6] and Singh et al. [9], patients with type 2 diabetes have a higher risk to develop dental caries. Also, Kampoo et al. [10] point out that the number of acidogenic bacteria in the mouth of diabetic patients is much higher than in healthy patients.

10. Feedback: In this step, the team considered to make further experiments to include a higher number of clinical records. To this end, the team came to the conclusion of the need for counting on a software tool to record clinical records to avoid the time-consuming process of digitalizing records in paper forms.

4) Conclusions and Future Work

This paper presented the application of data science to discover the relationship between dental caries and diabetes in dental records. Our results corroborate the relationship between diabetes and dental caries found in related work. This study opens an unexplored field in dentistry: the application of data science, based on a formal methodology and machine-learning techniques, to find hidden patterns in clinical records.

As future work, we are going to build a software tool to store and manage the clinical records of the Dental Clinic "Luz y Vida". The objective of this project is to reduce time when capturing and analyzing patients' data. Moreover, the features in the dataset will be extended with more features related to patients' lifestyle (e.g. exercise and nutrition habits). Also, we are going to carry out more experiments on this clinic's data about other diseases, which could be related to other dental pathologies.

References

[1] Sutthavong S, Sangasapaviriya A. Oral Health Care in Systemic Diseases. New York: Nova Science Publishers, Inc; 2009.

[2] Navarro AB, Faria R, Bascones A. Relación entre diabetes mellitus y enfermedad periodontal. Avances en Periodoncia. 2002;14(1):9-19.

[3] Juárez RP, Chahín, JR, Vizcaya MM, Arduña EI. Salud oral en pacientes con diabetes tipo 2: caries dental, enfermedad periodontal y pérdida dentaria. Odontología Sanmarquina. 2007;10(1):10-3.

[4] Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64-73.

[5] IBM. Foundational Methodology for Data Science. Available from https://public.dhe.ibm.com/common/ssi/ecm/im/en/imw14824usen/IMW14824USEN.PDF

[6] Seethalakshmi C, Reddy RCJ, Asifa N, Prabhu S. Correlation of Salivary pH, Incidence of Dental Caries and Periodontal Status in Diabetes Mellitus Patients: A Cross-sectional Study. Journal of Clinical

& Diagnostic Research. 2016;10(3):12-4.
[7] Novotna M, Podzimek S, Broukal Z, Lencova E, Duskova J. Periodontal Diseases and Dental Caries in Children with Type 1 Diabetes Mellitus. Mediators of Inflammation. 2015;2015:1-8.
[8] Miranda X, Troncoso J, Rodríguez C, Aravena P, Jiménez del R P. Caries e índice de higiene oral en niños con diabetes mellitus tipo 1. Revista Chilena de Pediatría. 2013;84(5):527-31.
[9] Singh I, Singh P, Singh A, Kour R. Diabetes an inducing factor for dental caries: A case control analysis. J Int Soc Prev Community Dent. 2016;6:125-9.
[10] Kampoo K, Teanpaisan R, Ledder RG, McBain AJ. Oral bacterial communities in individuals with type 2 diabetes who live in southern Thailand. Applied And Environmental Microbiology. 2014;80(2):662-71.
[11] Iqbal S, Kazmi F, Asad S, Mumtaz M, Khan AA. Dental caries & diabetes mellitus. Pakistan Oral & Dental Journal. 2011;31(1):58-61.
[12] Jawed M, Shahid SM, Qader SA, Azhar A. Dental caries in diabetes mellitus: role of salivary flow rate and minerals. Journal of Diabetes and its Complications. 2011;25(3):183-6.
[13] Miko S, Ambrus SJ, Sahafian S, Dinya E, Tamas G, Albrecht MG. Dental caries and adolescents with type 1 diabetes. British Dental Journal. 2010;208(6):E12-E.
[14] Stojanović N, Krunić J, Cicmil S, Vukotić O. Oral health status in patients with diabetes mellitus type 2 in relation to metabolic control of the disease. Srpski Arhiv Za Celokupno Lekarstvo. 2010;138(7-8):420-4.
[15] Hintao J, Teanpaisan R, Chongsuvivatwong V, Dahlen G, Rattarasarn C. Root surface and coronal caries in adults with type 2 diabetes mellitus. Community Dentistry And Oral Epidemiology. 2007;35(4):302-9.
[16] Camargo-Vega JJ, Camargo-Ortega, J. F., & Joyanes-Aguilar, L. Conociendo Big Data. Facultad de Ingeniería. 2015;24(38).
[17] Rivero D, Salgueiro-Silicia, Y, Domínguez, R. Valuation of several learning techniques in the Weka software. Innovación Tecnológica. 2012;18(3).
[18] Bogorny V, Avancini H, de Paula BC, Kuplich CR, Alvares LO. Weka-STPM: A Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. Transactions in GIS. 2011;15(2):227-48.
[19] Cambronero CG MI. Algoritmos de aprendizaje: knn & kmeans. Inteligencia en Redes Comun Univ Carlos III Madrid 2006.