Internship Outline

Student: Gatphan Atassi

Instructor: Dr. Jennifer Weller

Term: Summer 2020

## Introduction

The past 25 years have seen a rapid introduction and wide-spread implementation of diagnostic and cyber-information technologies in the medical and dental professions. These advancements have enabled more advanced detection and diagnosis of ailments in patients. However, these instruments and technologies can come at steep cost to the practice, hospital, and (inevitably) the patient. This is a result of new expensive machines and tools that require personnel training and heavy financial investment. It is this financial burden on the operators that gets passed to the patient.

In more recent years, there has been discussion and research into linking microbiomes and the presence (or lack thereof) of certain bacteria and viruses in fairly non-intrusive regions of the body. This marks a shift away from the previous approach of diagnosis wherein a photograph, radiographic image, or surface visibility were primary methods. Current initiatives focus on reducing patient discomfort and reducing intrusiveness of diagnosis methods while maintaining and increasing diagnosis accuracy.

My interest during this internship is to investigate, using 16S sequencing data, whether the ratios of bacteria present for a cavity are specific to the type of tooth. Additionally, it is hoped to find the certain concentrations of bacteria and viruses that exist when a carie is in the beginning stages of formation to aid in the detection of hidden caries in the early stages. My goal is to determine if it is possible to create a database of bacterial proportions wherein material from oral swaps can be sequenced and tested against this database in order to determine whether hidden caries are present in their early forms. Additionally, I aim to test whether caries in different regions of the mouth (which corresponds to different tooth types) will result in different proportions of bacterial/viral presence.

## Planned Method

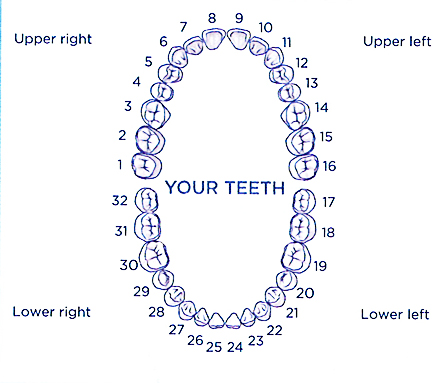
It is planned to construct a database wherein individuals are tracked anonymously through a casenum (case number) and a pid (patient identifier). The database will be organized in a star model-esque design. For exact specifications on the database construction, consult the schema below or the supplementary business rules. A few notes regarding these documents are to be addressed here, however. At this moment, gender is classified binarily. Future endeavors may expand the gender field to encompass non-binary people, but I believe this to beyond the scope of this experiment. Additionally, Each patient will have only 1 case number indicating their only visit wherein data is collected. While the database itself may handle cases of multiple cases per patient, such a scenario is not planned for at this moment. A patient will be tested during only one session but this session may render multiple data points in relation to the regions of the mouth tested. This eliminates the possibility of a patient dropping out of the study after only a single session and still provides the necessary point data needed.

Another detail to address is in regards to assay results. While the assay results may return more than 10 bacteria, the database table will only assign rank to the 10 most prevalent for retrieval simplicity and viewability. The raw data from the assays will preserve all the data in a separate table but will not be viewable in the database itself. If fewer than 10 bacterial species are reported, the name will be replaced with “NA”, while the percent presence will be set to -1. I do not expect such a case to occur but having a contingency for such a case is nevertheless useful.

Fungi such as Candida yeast will not be tracked in this study but may be added in future research.

With data collection like this, it is necessary for controls to be set up. That is why a control indicator. The Case number is an arbitrary integer value prepended with either ‘C’ or ‘E’. If a patient shows no disease symptom, they will be marked as a control and this will be reflected in their case number’s first letter being a ‘C’.

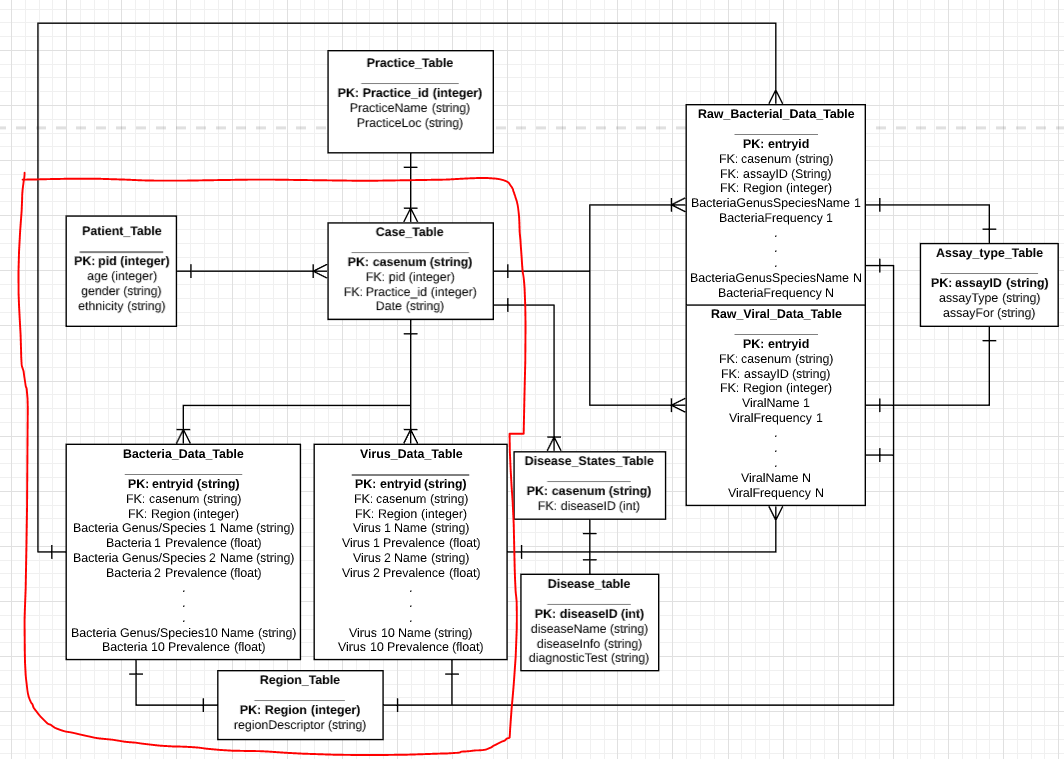
Organization of the oral cavity in this study will be slightly unorthodox. The human mouth contains four unique types of teeth: Incisors, canines, bicuspids (premolars), and molars. These are further distinguished by their location in the mouth. The further subdivisions include being maxillary (upper jaw) or mandibular (lower jaw) and left or right side of the mouth relative to the patient. These distinctions can be made for all teeth except for one case: the incisors which are only distinguished by whether they are maxillary or mandibular. In the dental profession adult teeth are numbered in a specific manner (see Figure 1). However, my goal is to focus on a regional investigation based on tooth type as the primary feature. As such, each subdivision will be numbered 1 through 14. The details are listed in the proof of concept data generation script (see appendix).



**Figure 1:** Numbering system of adult teeth

At this point one last clarification should be mentioned. Given the current COVID-19 pandemic, data acquisition may be an issue but a system of organization and analysis is possible to set up. To test the performance, synthetic data will be produced. It will have characteristics as similar to real data as feasible.

With the general format of the database discussed, the proposed database schema will appear somewhat like this:



**Figure 2: Database Schema where the tables circled in red denote those that will be simulated**

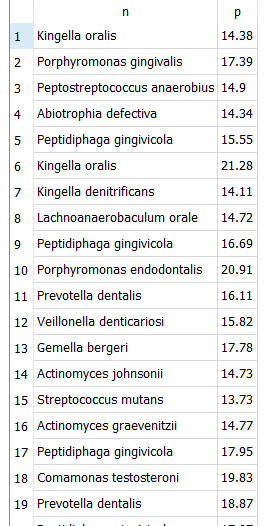
### Example Queries on the database

The following are a series of test queries that can be run on the mini-database created as a proof of concept. Note that I will be querying the bacterial data tables but all such commands can be easily replicated on it’s viral counterpart. Additionally, I will not be providing the entire output below the query commands but rather an excerpt of each

***Most common bacteria in Region 12 among our controls***

Select n,p From (Select Bacteria1Name as n, Bacteria1Prevalence as p, Bacteria\_Data\_Table.Region as r FROM Bacteria\_Data\_Table WHERE Bacteria\_Data\_Table.casenum like 'C%') WHERE r='12';

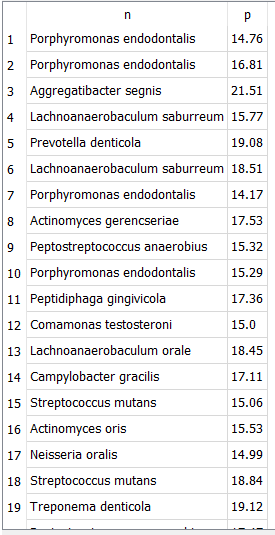
>



***Most common bacteria in Region 4 among the non-controls***

Select n,p From (Select Bacteria1Name as n, Bacteria1Prevalence as p, Bacteria\_Data\_Table.Region as r FROM Bacteria\_Data\_Table WHERE Bacteria\_Data\_Table.casenum like 'E%') WHERE r='4';

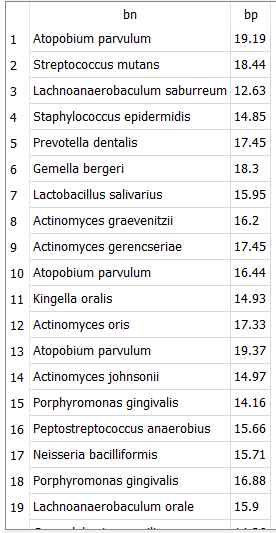
>



***Most common bacteria in region 5 among females***

SELECT bn, bp FROM( SELECT Bacteria1Name as bn, Bacteria1Prevalence as bp, Bacteria\_Data\_Table.Region as r From Bacteria\_Data\_Table INNER JOIN (Select Case\_Table.casenum as cn1 From Case\_Table INNER JOIN (SELECT PatientID as pid1 From Patient\_Table WHERE BiologicalGender like 'F%') ON Case\_Table.pid = pid1 ) ON Bacteria\_Data\_Table.casenum = cn1 ) WHERE r='5'

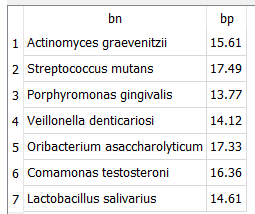
>



***Most common bacteria in region 11 among African Americans***

SELECT bn, bp FROM(SELECT Bacteria1Name as bn, Bacteria1Prevalence as bp, Bacteria\_Data\_Table.Region as r From Bacteria\_Data\_Table INNER JOIN (Select Case\_Table.casenum as cn1 From Case\_Table INNER JOIN (SELECT PatientID as pid1 From Patient\_Table WHERE Patient\_Table.ethnicity like 'African%') ON Case\_Table.pid = pid1) ON Bacteria\_Data\_Table.casenum = cn1) WHERE r='11'

>



***Count of Males who are controls (no diseases) compared to Count of Females who are controls***

SELECT COUNT (g) FROM (SELECT g,cn FROM (SELECT Patient\_Table.BiologicalGender as g, Case\_Table.casenum as cn FROM Case\_Table INNER JOIN Patient\_Table ON Case\_Table.pid = Patient\_Table.PatientID) WHERE g like 'M%');

>56

SELECT COUNT (g) FROM (SELECT g,cn FROM (SELECT Patient\_Table.BiologicalGender as g, Case\_Table.casenum as cn FROM Case\_Table INNER JOIN Patient\_Table ON Case\_Table.pid = Patient\_Table.PatientID) WHERE g like 'F%');

>44

### Summary

Although the data used in this internship was procedurally generated as a proof of concept there are some valuable takeaways to be had. Firstly, when looking at the data tables and the queries I can push through the database, there are some fairly powerful conclusions that could be made if real data was collected/used. Being able to collate gender, race, age, oral region, disease state, etc to analyze populations prevalence has the potential to open the door towards the expansion of preventative oral care and perhaps new diagnostic methods for whole body diseases. While no conclusions like that can be made with this project, what can be said is that the potential is clearly there to be expanded upon. There are some weaknesses however. Firstly, this tool can only be as powerful as the number of patients analyzed (which in itself presents personal information concerns that would need to be addressed). There is not a concrete number of patients that is currently known that would render this database as a viable tool. Additionally, this database would rely heavily on the strength of assays to identify taxa properly. This in itself adds another layer of potential weakness. Furthermore, this database has room to expand. Perhaps adding a geography field denoting where the data was collected. Additionally adding the nonbinary gender spectrum may be good route of expansion along side the current biological gender. All in all, I believe that there is a road where this project may deliver useful if not powerful data analysis but not without addressing some pertinent concerns.