Principal Component Analysis observing different parameters of medical factors influencing dental health conditions of pediatric patients

by

Gabrielle Ulery

Bachelor's in Science in Biological Sciences, University of Maryland, 2019

Submitted to the Graduate Faculty of the

School of Dental Medicine in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF DENTAL MEDICINE

This thesis was presented

by

Gabrielle Ulery

It was defended on

April 13, 2022

and approved by

Alexandre R Vieira, DDS MS PhD, Oral and Craniofacial Sciences

Juan Taboas, MS PhD, Oral and Craniofacial Sciences

Adriana Modesto Vieira, DDS DMD MS PhD, Department of Pediatric Dentistry

Copyright © by Gabrielle Ulery

2022

Principal Component Analysis observing different parameters of medical factors influencing dental health conditions of pediatric patients

Gabrielle Ulery, MS

University of Pittsburgh, 2022

Pediatric patients are in comparison extremely different than adult patients. They require different types of care and more preventative oral health procedures. While this is true, it is important to understand variables and factors that could correlate with their dental health conditions and treatments. In this research experiment, we determined different parameters from pediatric patients medical and dental health evaluations to see if there is any similarity in specific medical factors influencing dental health conditions. Data from 528 pediatric subjects 0-18 of age were obtained from the Dental Registry and DNA Repository project. Data included demographics, self-reported medical history, medications taken, and oral health conditions. Principal component analysis (PCA) was used to cluster similar individuals aiming to generate homogeneous groups for future genetic studies. We found that patients were more similar in comparison than initially predicted. However, there was a trend suggesting there were two groups. These groups were further investigated by using single nucleotide polymorphisms marking *TRAV4* and *MMP2*. There were no differences in the distribution of the SNP markers between the two groups. We cannot further explain why the two groups may differ, and future studies will aim to explain the PCA initial findings.

Table of Contents

List of Tables

List of Figures

1.0 Background Information

Pediatric dental patients are extremely different than adults. Children are actively growing and developing. Throughout previous research, children have proven to have different clinical presentations and trajectory of disease diagnosis (Vieira 2020). When looking at oral, dental, and craniofacial conditions, children have their own very distinct features, in comparison to the same conditions that affect adults differently (Vieira 2020). In different research findings, these conditions can affect adults will not apply to children (Vieira 2020). Pediatric patients need to be the central focus of research on the definition of what is best for their care. There are many gaps of knowledge in pediatric dental care that need to be assessed and determined (Vieira 2020). They require different types of care and more preventative oral health procedures (Vieira 2020). It is important to understand variables and factors that could correlate with their dental health conditions and treatments. It is important to see how medical conditions and factors can be related to dental factors in children as well. If we can establish a relationship between the two, we can have better knowledge to help children and explore different trajectory of disease diagnosis and help children.

Not all pediatric patients have the same access to better oral health resources when compared to others. In the United States of America, the nation's oral health has greatly improved since the 1960s, but not all Americans have equal access to these improvements (National Institute of Dental and Craniofacial Research 2000). Some individuals and racial/ethnic groups have worse oral health due to the result of social determinants of health (Braveman et al. 2011). Many people are unable to afford dental care and maintain regular preventative dental care that is essential for good oral health (CDC 2022).

It is important to understand the profile of the children in the region being tested in this study. The current profile of children is from Pittsburgh, USA. Pittsburgh is the largest city of the poorest area in the country, the Appalachian region (Vieira et al. 2020). This area is known to have some of the worse health outcomes of the country (Vieira et al. 2020). Many aspects of the region might not be generalizable, but also an area to develop new studies (Vieira et al. 2020). This study uses data from the Dental Registry and DNA Repository project (DRDR) at the University of Pittsburgh School of Dental Medicine. The DRDR currently has more than 6,700 participants and individuals provide written consent to allow or their medical and dental records to be used for future investigations and give a biological sample (saliva) that allows for the planning of molecular sciences (Vieira et al. 2020).

1.1 Principal Component Analysis

This information can be observed through a statistical test called principal component analysis. Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by multiple inter-correlated quantitative dependent variables (Abdi and Williams 2020). PCA analyzes a data table representing observations that are described by several dependent variables that are inter-correlated and allows the extraction of important information from the table to represent it as a new set of orthogonal variables called principal components, to display the pattern of similarity in the observations and of the variables in certain locations on maps (Abdi and Williams 2020).

Principal component analysis (PCA) is a useful technique for exploratory data analysis, allowing better visualization in the variation present in a dataset that contains many variables

(Hayden 2018). PCA allows the overall 'shape' of the data to be seen and identifies which samples are like one another and which are very different from one another (Hayden 2018). By using PCA, it can enable the identity of groups of samples that are similar and establish which variables make a group different from another (Hayden 2018). The intended goals of PCA are

(1) extract the most important information from the data table,

(2) compress the size of the data set by only keeping this important information,

(3) simplify the description of the data set; and

(4) analyze the structure of the observations and the variables

(Abdi and Williams 2020).

The variables that were observed in this study included demographics, self-reported medical history, medications taken, and oral health conditions. We can characterize similarities and differences in these variables and the patients by using principal component analysis. This allows the ability to cluster similar individuals aiming to generate homogenous groups for future genetic studies. In this research experiment, we determined different parameters from pediatric patients medical and dental health evaluations to see if there is any similarity in specific medical factors influencing dental health conditions.

1.1.1 T-Cell Receptor Alpha Chain Variable 4 (*TRAV4*)

Caries is still a major problem affecting 60 to 90% of children (World Health Organization 2003). The etiology is complex and multifactorial with contributions from external factors from the host, the type of diet, and practices and habits within the family, and pressures from society (Fisher-Owens et al. 2007). The way to identify at risk children are through classic methods which include application of diet questionnaires, inspection of oral hygiene level, and detection of

Streptococcus mutans in saliva (Briseño-Ruiz et al. 2014). These methods are limited in determining caries risk at population level (Briseño-Ruiz et al. 2014).

There is different expression in genes and proteins in whole saliva that can help identify risk for caries. Improved understanding of relevant genetic factors will help increase the precision of caries risk assessment and will also enable more targeted approaches that will prevent and manage dental caries (Wright 2019). For example, in chromosome 14, genetic markers flanking T-cell Receptor Alpha Chain Variable 4 (*TRAV4)* were associated with low caries experience and *TRAV4* expression in the whole saliva of individuals with low caries experience was higher in children and teenagers when being compared to adults (Vieira et al. 2014). Variations in the salivary protein T-cell Receptor Alpha Chain Variable 4, and the gene that codes for this protein (*TRAV4*) is associated with low caries experience (Wright 2019). The fine mapping of the locus 14q11.2 showed *TRAV4* as involved in caries experience (Briseño-Ruiz et al. 2014).

Previous caries experience continues to be the best predictor for future disease (Powell 1998). A child's resistance or susceptibility to caries can occur, regardless to exposure of external risk factors (Briseño-Ruiz et al. 2014). These facts suggest there is a biological influence on disease susceptibility which is likely to be controlled by genetic factors of the host (Briseño-Ruiz et al. 2014). There have been studies that have previously indicated that there is a higher expression of the gene in children and teenagers with low caries experience, correlating with specific alleles in *TRAV4* (Briseño-Ruiz et al. 2014). They assessed the mRNA expression of *TRAV4* in the saliva of 143 study subjects in from Argentinian families and tested their statistically significant associations that was found between low caries experience and markers in *TRAV4* (Briseno-Ruiz et al. 2014). They were able to replicate the initial genetic association results in additional populations that were characteristically from underserved areas (Briseño-Ruiz et al. 2014). Their

results suggested that *TRAV4* may have a role in protecting against caries (Briseño-Ruiz et al. 2014). *TRAV4* was chosen in this study due to previous research studies to establish a role in restorations from PCA findings of pediatric population.

1.1.1.1 Matrix Metalloproteinase Protein 2 (*MMP2*)

Another gene family that has been associated with caries is the matrix metalloproteinases (*MMPs*). The matrix metalloproteinase genes (*MMPs*) are fundamental in the tooth formation and mineralization of dental tissue in rats during the formation of enamel and dentin (Fanchon *et al*. 2004). Matrix metalloproteinases and their inhibitors might be involved in enamel formation (Antunes et al. 2015). Matrix metalloproteinases (*MMPs*) are associated with levels of inflammation and are involved in caries, pulpal, and periapical tissue destruction (Menezes-Silva et al. 2012). *MMPs* also play a huge role in bone resorption and in previous finding have polymorphisms in *MMP* genes and their regulators may contribute to an individual's increased susceptibility to apical tissue destruction in responses to deep carious lesions (Menezes-Silva et al. 2012). Other studies demonstrated that genetic variants in *MMPs* might be involved in caries susceptibility.

Matrix metalloproteinases play an important role during the initial process of enamel development and therefore may play a role in caries (Tannure et al. 2012). MMPs are associated with levels of inflammation and are involved in caries (Menezes-Silva et al. 2012). There has been genetic variation in *MMP* genes that influence progression of carious lesions in dentin and development of periapical pathology, since *MMPs* are involved with dentin and bone degradation (Menezes-Silva et al. 2012). Human genetic polymorphisms appear to play a role in the disease susceptibility of the host (Menezes-Silva et al. 2012). They found that the combined bacterial/host

genotyping may be providing an important tool in the definition of disease risk and targeting bacteria eradication to high-risk individuals (Menezes-Silva et al. 2012).

Since it is known that *MMPs* impact the progression of caries lesion into the dentin, variants in *MMP2* and *MMP3* have been associated in periapical lesion formation (Menezes-Silva et al. 2012). MMP2 is a gelatinase involved in mineralization and dentin degradation and catalyzes dentin matrix degradation after mineralization (Niu et al. 2011). There have also been studies done discussing the role of *MMP2* in failure of restorations due to secondary caries (Benli et al. 2021). The *MMP2* variation impacts the risk of having secondary caries, independent of the restorative material (Benli et al. 2021). *MMP2* has been associated with failure of dental restorative treatments and have the potential to be used for determination of risks impacting longevity of dental treatments (Benli et al. 2021). *MMP2* was chosen from previous research findings in this study to assess the restorations from the PCA graph.

2.0 Aims

The aims of this Master's thesis were to

2.1 Aim 1 Define the Pediatric Population Sample from University of Pittsburgh School of Dental Medicine from the Dental Registry and DNA Repository Project (DRDR)

Aim 1 established the sample that was chosen from the University of Pittsburgh School of Dental Medicine from the Dental Registry and DNA Repository Project. The sample size was 528 pediatric subjects 0-18 years of age.

2.2 Aim 2 Principal Component Analysis of Pediatric Subjects using RStudio

Aim 2 uses the medical factors and dental health conditions noted by the pediatric subjects in the DRDR as variables. These variables were translated into a numerical coding system, which allowed the use of the statistical software RStudio. The variables were compared in a biplot showing similarity to one another.

2.3 Aim 3 Test for Association between the Two Groups obtained from the PCA and variants in *TRAV4* and *MMP2*.

Aim 3 uses the separation in pediatric subjects in divisions due to location on biplot. The left split is labeled “Northwest” and the right split is labeled “Northeast”. Pediatric subjects were identified, and their saliva samples was used to extract genomic DNA and determine genotypes to test for *TRAV4* and *MMP2* association with the groups using SNPs rs1997532, rs1997533, rs1784418, and rs7150049.

In this research experiment, we hypothesize that there will be an overall similarity within the pediatric population of their specific medical factors influencing dental health conditions that can be used to generate homogeneous groups for future genetic studies.

3.0 Materials and Methods

3.1 Sample Population Selection

The data used in this study was obtained by from 528 pediatric subjects 0-18 of age from the University of Pittsburgh School of Dental Medicine Dental Registry and DNA Repository project (DRDR) (Vieira 2020). The pediatric subjects were asked for consent to obtain a saliva sample that will be stored for future studies (Vieira 2020). The patients' medical history and dental records were exported into an Excel sheet and variables for PCA were selected. The variables that were being tested were sex, ethnicity, restorations, oral infections, oral surgery, root canal, periodontics, prosthodontics, orthodontics, oral prophylaxis, general health state, if the participant had been hospitalized in the last 5 years, if the participant was under a physicians care, if the patient took any medication, if the patient had have allergies, drug use, alcohol use, if the participant currently used tobacco, if the participant ever used tobacco, history of medical use, epilepsy, fainting, stroke, asthma, anemia, sinus problems, hepatitis, bruise or bleed easily, irregular heartbeat, high blood pressure, mitral valve prolapse, heart murmur, congenital heart lesions, artificial heart valves, cancer, cancer radiation, heart surgery, kidney disease or dialysis, diabetes, prosthetic joints, and HIV AIDS.

3.2 Translation for RStudio from Variables for Principal Component Analysis

Data were then converted into a codable sheet in Excel for the software, RStudio. The variables from the patient's medical history and dental records were converted into a numerical code. RStudio can use the numerical order for PCA to separate the variables based on comparability.

Gender:
Male- 1
Female- 2
Decline answer or Unknown: 3
Ethnicity:
Caucasians- 1
African- Americans- 2
Asians- 3
Hispanic- 4
Other- 5
Restorations:
Yes- 1
No-2
N/A- 3
Oral Infections:
Yes- 1
No-2
N/A- 3
Oral Surgery:
Yes- 1
No-2
N/A- 3
Root Canals:
Yes- 1
No-2
N/A- 3
Periodontal:
Yes- 1
No-2
N/A- 3
Prosthodontics:
Yes- 1
No-2
N/A- 3

Orthodontics:
Yes- 1
No-2
N/A- 3
Oral Prophylaxis:
Yes- 1
No-2
N/A- 3
General Health State:
Excellent- 1
Good- 2
Fair- 3
N/A- 4
Has the participant been hospitalized in the last 5 years?
Yes- 1
No- 2
N/A- 3
Is the participant under a physician's care?
Yes- 1
No- 2
N/A- 3
Does the patient take medications?
Yes- 1
No- 2
N/A- 3
Does the patient have allergies?
Yes- 1
No- 2
N/A- 3

Drug Use:
Yes- 1
No- 2
N/A- 3
Alcohol Use:
Yes- 1
No- 2
N/A- 3
Tobacco Use:
Yes- 1
No- 2
N/A- 3
History of mental health issue:
Yes- 1
No-2
N/A- 3
Epilepsy
Yes- 1
No- 2
N/A- 3
Fainting
Yes-1
No-2
N/A- 3
Stroke
Yes-1
No-2
N/A- 3
Asthma
Yes- 1
No- 2
N/A-3
Sinus Problems
Yes- 1
No- 2
N/A- 3

Hepatitis
Yes- 1
No- 2
N/A- 3
Bruise or Bleed Easily
Yes- 1
No- 2
N/A- 3
Irregular Heartbeat
Yes- 1
No- 2
N/A- 3
High Blood Pressure
Yes- 1
No- 2
N/A- 3
Mitral Valve Prolapse
Yes- 1
No- 2
N/A- 3
Congenital Heart lesions
Yes- 1
No- 2
N/A- 3
Artificial Heart Valves
Yes- 1
No- 2
N/A- 3
Cancer
Yes- 1
No- 2
N/A- 3
Cancer Radiation
Yes- 1
No- 2
N/A- 3

Heart Surgery
Yes- 1
No- 2
N/A- 3
Kidney Disease or Dialysis
Yes- 1
No- 2
N/A- 3
Diabetes
Yes- 1
No- 2
N/A- 3
Prosthetic Joints
Yes- 1
No- 2
N/A- 3
HIV AIDS
Yes- 1
No- 2
N/A- 3

Figure 1. Coding Parameters of Medical History and Dental Records

3.3 RStudio Principal Component Analysis Coding Process

A specific coding outline was followed using a RTutorial on Datacamp.

```
mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE,scale. = TRUE)

summary(mtcars.pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.3782 1.4429 0.71008 0.51481 0.42797 0.35184
## Proportion of Variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375
## Cumulative Proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936
##                            PC7    PC8     PC9
## Standard deviation     0.32413 0.2419 0.14896
## Proportion of Variance 0.01167 0.0065 0.00247
## Cumulative Proportion  0.99103 0.9975 1.00000
```

Figure 2. Coding Template from Datacamp

```
str(mtcars.pca)

## List of 5
##  $ sdev    : num [1:9] 2.378 1.443 0.71 0.515 0.428 ...
##  $ rotation: num [1:9, 1:9] -0.393 0.403 0.397 0.367 -0.312 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:9] "mpg" "cyl" "disp" "hp" ...
##   .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:9] 20.09 6.19 230.72 146.69 3.6 ...
##   ..- attr(*, "names")= chr [1:9] "mpg" "cyl" "disp" "hp" ...
##  $ scale   : Named num [1:9] 6.027 1.786 123.939 68.563 0.535 ...
##   ..- attr(*, "names")= chr [1:9] "mpg" "cyl" "disp" "hp" ...
##  $ x       : num [1:32, 1:9] -0.664 -0.637 -2.3 -0.215 1.587 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##   .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"
```

Figure 3. Coding Key for PCA

The dataset uses 38 medical history and dental records, taken from DRDR. The matrix of 9 columns and 38 rows was assigned **prcomp()** function, assigning output **RESEARCHnew.pca**.

Two arguments, **center** and **scale**, were set to be **TRUE**. Then PCA object was obtained with **summary()** (Hayden 2018).

After following a specific coding series in RStudio, a biplot was made. This was repeated multiple times to ensure efficiency.

3.4 Allelic Discrimination

There was an apparent split in the pediatric patients from the PCA biplot. After identifying the patients, they were separated into two groups. The genotyping of markers, *TRAV4* and *MMP2*, were chosen based on the separation in the data being related to sex and oral infections. The genotyping of markers in *TRAV4* and *MMP2* was obtained for 157 patients of the "Northwest" and the Northeast side of the pediatric subjects included 84 patient's region from previous graphs due to the location found using RStudio of the patients and the amount of saliva in the remaining sample available. The concentrations of genomic DNA obtained from the samples of each pediatric patient from Northwest and Northeast portions were obtained using a spectrophotometer and dilutions to 2ng/µl were calculated in Excel to be diluted with buffer 1XTE Buffer made 6/19/20 and are shown in Table 1.

Table 1. Concentrations of Genomic DNA of Northwest Group with Buffer Concentrations

Participant #	Concentration from DRDR	Sample Concentration	Buffer Concentration
20	108.21	1.848258	98.15174
24	153.37	1.304036	98.69596
31	158.5	1.26183	98.73817
39	121.505	1.646023	98.35398
45	66.64	3.0012	96.9988

66	105.445	1.896723	98.10328
67	145.56	1.374004	98.626
69	80.295	2.490815	97.50918
72	101.17	1.976871	98.02313
74	75.05	2.66489	97.33511
83	73.17	2.733361	97.26664
87	319.58	0.625821	99.37418
91	121.28	1.649077	98.35092
97	67.955	2.943124	97.05688
100	136.98	1.460067	98.53993
105	142.61	1.402426	98.59757
108	5.3	37.73585	62.26415
118	100.415	1.991734	98.00827
119	100.5	1.99005	98.00995
124	78.88	2.535497	97.4645
148	114.48	1.74703	98.25297
161	92.88	2.153316	97.84668
170	37.37	5.351887	94.64811
176	97.38	2.05381	97.94619
181	65.11	3.071725	96.92828
194	278.92	0.717051	99.28295
202	153.2	1.305483	98.69452
204	142.69	1.40164	98.59836
206	90.05	2.220988	97.77901
210	131.77	1.517796	98.4822
212	129.42	1.545356	98.45464
214	135.18	1.479509	98.52049
222	31.395	6.370441	93.62956
223	2.03	98.52217	1.477833
238	120.195	1.663963	98.33604
248	55.66	3.593245	96.40676
249	49.895	4.008418	95.99158
251	143.415	1.394554	98.60545
297	180.36	1.108893	98.89111
275	70.74	2.827255	97.17275
282	31.6	6.329114	93.67089
291	95.11	2.102828	97.89717
298	23.49	8.514261	91.48574
310	81.535	2.452934	97.54707
312	110.515	1.809709	98.19029
315	23.54	8.496177	91.50382
317	98.505	2.030354	97.96965
325	19.06	10.49318	89.50682
335	96.075	2.081707	97.91829
341	73.045	2.738038	97.26196

348	130.025	1.538166	98.46183
373	128.19	1.560184	98.43982
388	124.44	1.6072	98.3928
389	181.85	1.099808	98.90019
390	177.44	1.1127142	98.87286
392	95.64	2.091175	97.90882
397	144.825	1.380977	98.61902
406	123.355	1.621337	98.37866
408	68.81	2.906554	97.09345
410	131.76	1.517911	98.48209
421	84.61	2.363787	97.63621
425	24.88	8.038585	91.96141
427	28.99	6.898931	93.10107
436	81.295	2.460176	97.53982
438	39.85	5.018821	94.98118
447	146.66	1.363698	98.6363
448	137.585	1.453647	98.54635
465	59.05	3.38696	96.61304
470	86.575	2.310136	97.68986
471	16.97	11.7855	88.2145
479	122.075	1.6383337	98.36166
481	139.6	1.432665	98.56734
489	113.21	1.766628	98.23337
692	62.8	3.184713	96.81529
733	32.9	6.079027	93.92097
763	165.6	1.207729	98.79227
783	42.24	4.734848	95.26515
910	132.03	1.514807	98.48519
921	56.82	3.519887	98.48011
1008	31.25	6.4	93.6
1058	9.77	20.47083	79.52917
1245	64.04	3.123048	96.87695
1266	148.51	1.346711	98.65329
1357	35.03	5.709392	94.29061
1372	32.38	6.176652	93.82335
1384	101.63	1.967923	98.03208
1401	33.87	5.904931	94.09507
1540	24.78	8.071025	91.92897
1731	24.27	8.240626	91.92897
1741	5.57	35.90664	64.09336
1836	34.42	5.810575	94.18942
1935	182.12	1.098177	98.90182
1939	138.66	1.442377	98.55762
1951	208.25	0.960384	99.03962
1953	277.34	0.721137	99.27886

1962	27.25	7.33945	92.66055
1968	5.8	34.48276	65.51724
2017	150.69	1.327228	98.67277
2019	258.65	0.773246	99.22675
2113	112.53	1.777304	98.2227
2114	176.79	1.131286	98.86871
2153	238.16	0.839772	99.16023
2366	128.88	1.551831	98.44817
2673	211.02	0.947777	99.05222
2708	216.51	0.923745	99.07626
2713	14.4	13.88889	86.11111
2721	33.94	5.892752	94.10725
2722	33.07	6.047777	93.95222
2724	6.13	32.62643	67.37357
2728	118.77	1.683927	98.31607
2729	2.34	85.47009	14.52991
2749	114.29	1.749934	98.25007
2761	73.05	2.737851	97.26215
2764	64.94	3.079766	96.92023
2771	55.05	3.633061	96.36694
2779	112.8	1.77305	98.22695
2782	72.04	2.776235	97.22376
2976	32.11	6.228589	93.77141
3018	135.19	1.479399	98.5206
3030	398.31	0.502121	99.49788
3117	100.89	1.982357	98.01764
3152	299.51	0.667757	99.33224
3499	105.7	1.892148	98.10785
3500	61.1	3.273322	96.72668
3551	50.85	3.933137	96.06686
3552	88.88	2.250225	97.74977
3601	100.39	1.99223	98.00777
3900	238.97	0.836925	99.16307
4658	81.5	2.5	97.5
4821	2117.1	0.094469	99.90553
4825	333.7	0.6	99.4
4839	173.8	1.2	98.8
4946	114.4	1.7	98.3
4973	94.3	2.1	97.9
4981	55.8	3.6	96.4
5252	8.0	25.1	74.9
5612	25.0	8.0	92.0
5791	27.4	7.3	92.7
5830	131.5	1.5	98.5
6023	50.6	3.9	96.1

6358	20.4	9.8	90.2
6379	46.6	4.3	95.2
6395	42.8	4.7	95.3
6397	44.5	4.5	95.5
6412	19.3	10.3	89.7
6414	38.3	5.2	94.8

Table 2. Concentrations of Genomic DNA of Northeast Group with Buffer Concentrations

Participant #	Concentration from DRDR	Sample Concentration	Buffer Concentration
692	62.8	3.2	96.8
733	32.9	6.1	93.9
763	165.6	1.2	98.8
783	42.2	4.7	95.3
910	132.0	1.5	98.5
921	56.8	3.5	96.5
1008	31.3	6.4	93.6
1058	9.8	20.5	79.5
1245	64.0	3.1	96.9
1266	148.5	1.3	98.7
1357	35.0	5.7	94.3
1372	32.4	6.2	93.8
1384	101.6	2.0	98.0
1401	33.9	5.9	94.1
1540	24.8	8.1	91.9
1731	24.3	8.2	91.8
1741	5.6	35.9	64.1
1836	34.4	5.8	94.2
1935	182.1	1.1	98.9
1939	138.7	1.4	98.6
1951	208.3	1.0	99.0
1953	277.3	0.7	99.3
1962	27.3	7.3	92.7
1968	5.8	34.5	65.5
2017	150.7	1.3	98.7
2019	258.7	0.8	99.2
2113	112.5	1.8	98.2
2114	176.8	1.1	98.9
2153	238.2	0.8	99.2
2366	213.0	0.9	99.1
2673	211.0	0.9	99.1

2708	216.5	0.9	99.1
2713	14.4	13.9	86.1
2721	33.9	5.9	94.1
2722	33.1	6.0	94.0
2724	6.1	32.6	67.4
2728	118.8	1.7	98.3
2729	2.3	100.0	0.0
2749	114.3	1.7	98.3
2761	73.1	2.7	97.3
2764	64.9	3.1	96.9
2771	55.1	3.6	96.4
2779	112.8	1.8	98.2
2782	72.0	2.8	97.2
2976	32.1	6.2	93.8
3018	135.2	1.5	98.5
3030	398.3	0.5	99.5
3117	100.9	2.0	98.0
3152	299.5	0.7	99.3
3499	105.7	1.9	98.1
3500	61.1	3.3	96.7
3551	50.9	3.9	96.1
3552	88.9	2.3	97.7
3601	100.4	2.0	98.0
3900	239.0	0.8	99.2
4342	167.1	1.2	98.8
4658	81.5	2.5	97.5
4720	294.7	0.7	99.3
4820	337.7	0.6	99.4
4821	834.6	0.2	99.8
4825	333.7	0.6	99.4
4839	173.8	1.2	98.8
4889	85.7	2.3	97.7
4931	26.2	7.6	92.4
4946	114.4	1.7	98.3
4973	94.3	2.1	97.9
4981	55.8	3.6	96.4
5252	8.0	25.0	75.0
5612	25.0	8.0	92.0
5791	27.4	7.3	92.7
5830	131.5	1.5	98.5
6023	50.6	4.0	96.0
6358	20.4	9.8	90.2
6379	46.6	4.3	95.7
6395	42.3	4.7	95.3
6397	44.5	4.5	95.5

6412	19.3	10.4	89.6
6414	38.3	5.2	94.8

The patients DNA sample was diluted to 2ng/µl in a 96 well plate. The first Reaction Mix was made based on the following formula using *TRAV4* and *MMP2* single polymorphism nucleotide (SNP) probes, allowing at least four extra wells for a negative control and general loss. The *TRAV4* SNPs that were included in the Reaction Mix were *rs1997532, rs7150049,* and *rs1997533*. The *MMP2* SNP that was included was *rs1784418*. The Reaction mix included: Master Mix- 1.5 µl, 40X SNP- 0.037 µl, Water- 0.462 µl, total- 2 µl per well. Added 1 µl of diluted DNA from the 96 well plate to the 384 well plate with the multi-channel pipettor. Added 2 µl of the reaction mix to each well with the electronic repeater pipettor. One well had 2 µl of mix, but no DNA. Water was added, negative control obtained. Adhesive was pressed onto the plate using plastic spatula, plate was not touched, and film was centered on the plate. Ran plate in a thermocycler on the program: 95 °C for 10 minutes, 40 cycles of [92°C for 15 seconds and 60°C for 1 minute].

After amplification, plate was then placed in **QuantStudio** 6 Flex machine and run using **QuantStudio™** Real-Time PCR System for allelic discrimination. The experiment was set up to run **QuantStudio™** Flex System with a 384 well for genotyping using **TaqMan**® reagents for a standard run. The SNP Assay were named with plate layout. Run method was selected and genotypes were found and shown via Amplification Plot.

3.5 Statistical Summary of Genotypes

Chi-square was used to test for Hardy-Weinberg equilibrium. P-values below 0.001 were considered to be not in equilibrium. Chi-square was used for testing over-representation of genotypes or alleles between the two groups with an alpha of 0.05.

4.0 Results

Biplots were used to display the results of the PCA when comparing the variables tested of pediatric patients. The dots represent patients, while the arrows demonstrate the variables being tested. When running principal component analysis, the closer the variables are, the more similar the patients are to that variable. A biplot is a type of plot that allows visualization on how the samples relate to one another in PCA and shows which samples are similar and which are different. This will simultaneously reveal how each variable contributes to each principal component (Hayden 2018). The axes are originating from the center point. The variables previously mentioned contribute to PC1, with higher values in those variables moving the samples to the right of the plot. The data points relate to the axes and will be further investigated in identifying patients. PCA compares the parameters from coding key in R. The closer the parameters are together, the more comparability there is.

In Figure 4, the 528 patients are observed in biplot created in RStudio. Patients were compared to variables order of participant (order), sex, ethnicity, restorations (R), oral infections (OI), oral surgery (OS), root canals (RC), periodontal (P), prosthodontics (PR), orthodontic, oral prophylaxis (OP), general health state (H), is the participant under a physician care? (PC), drug use (D), alcohol (A), and history of mental health (MH). Sex and Ethnicity arrows were close together, demonstrating variable similarities. Majority of other procedure arrows were close together or overlapping, demonstrating more similarities in pediatric participants in the population than expected. Outliers on right side were further investigated.

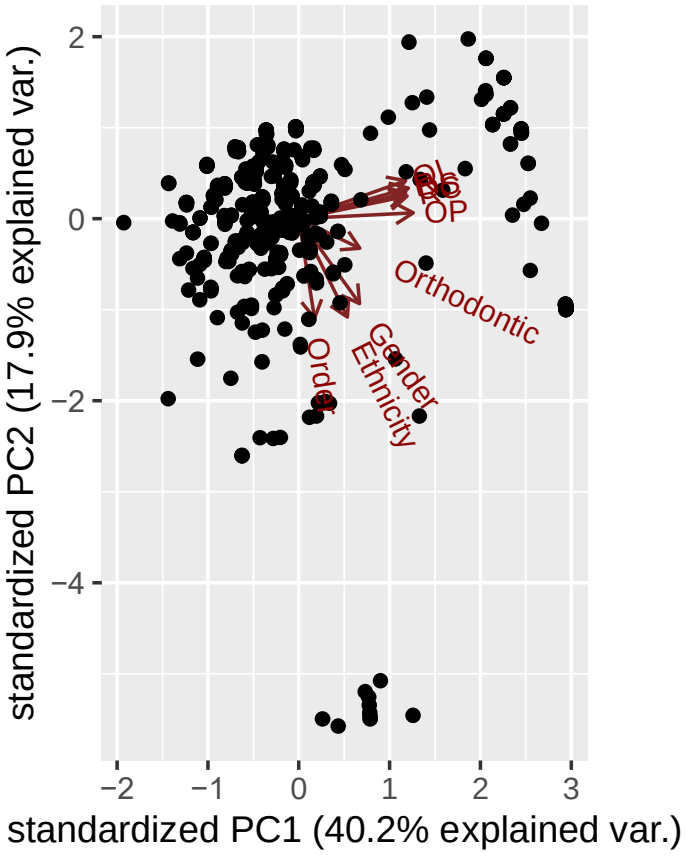

Figure 4. 528 pediatric patients were observed in a PCA biplot using RStudio. There were 16 variables observed between the patient's medical history and dental records. Sex and ethnicity were in proximity, representing comparability. Other dental records and medical history arrows were close together or overlapping, demonstrating more similarities in pediatric participants in the population than expected.

Figure 5 is continuation of observation from Figure 4, but different sizing scale. This establishes a different visual perspective in analyzing the parameters from the PCA run. Patients of particular interest were outliers on right side of biplot.

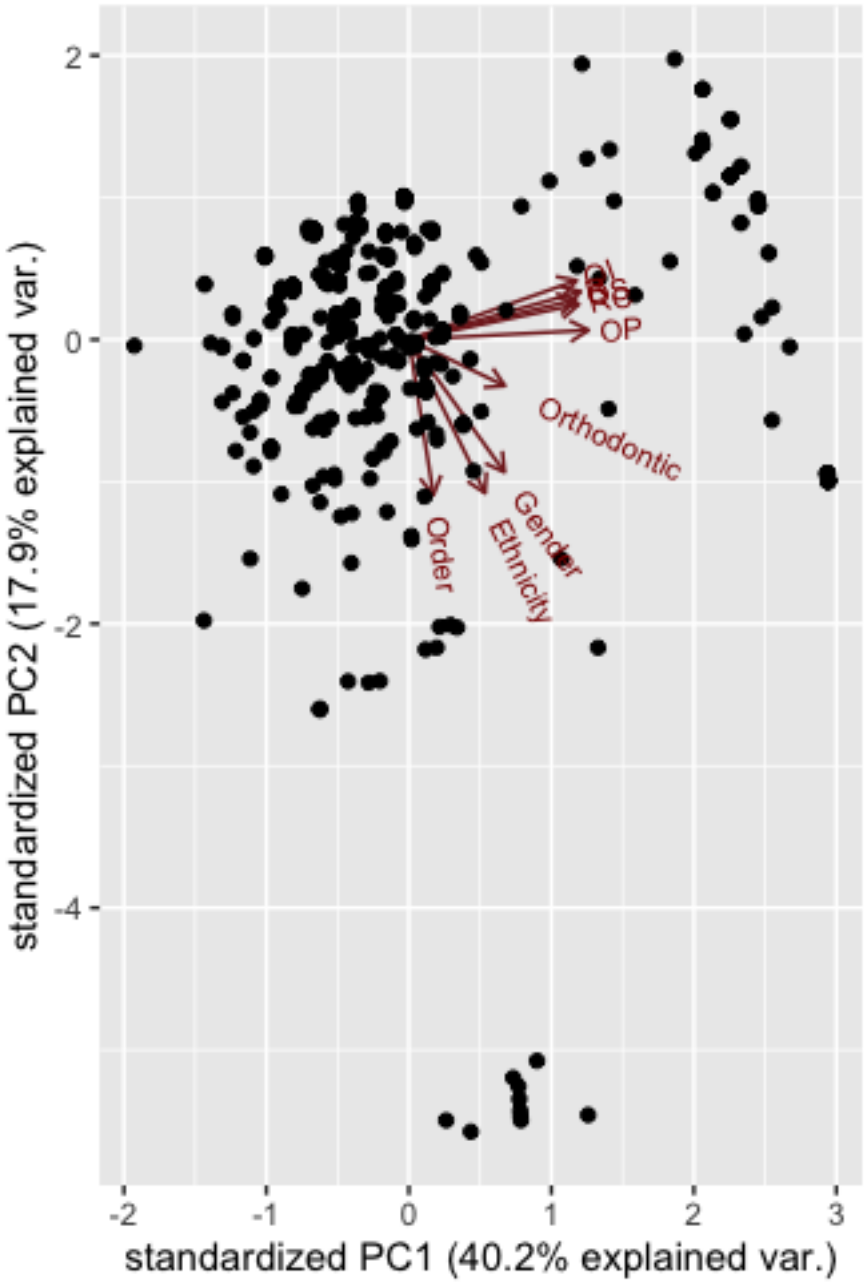

Figure 5. Continuation of observation from Figure 1, but different sizing scale. This establishes a different visual perspective in analyzing the variables from the PCA biplot. Patients of particular interest were outliers on right side of biplot, but are ultimately excluded from the experiment.

After finding similarities between parameters and participants in the biplots, the participants needed to be identified to further investigate the relationship. Figure 6 displays the

order of participated was converted to numbers of data to R. This allowed the ability to easily identify patients when coding in RStudio. By identifying patients in their selected order numbers, their dental and medical histories could be analyzed. This will help distinguish and establish which patients have variables in comparison and what makes them unique to the study. Participants of particular interest are outliers found on right side of biplot.

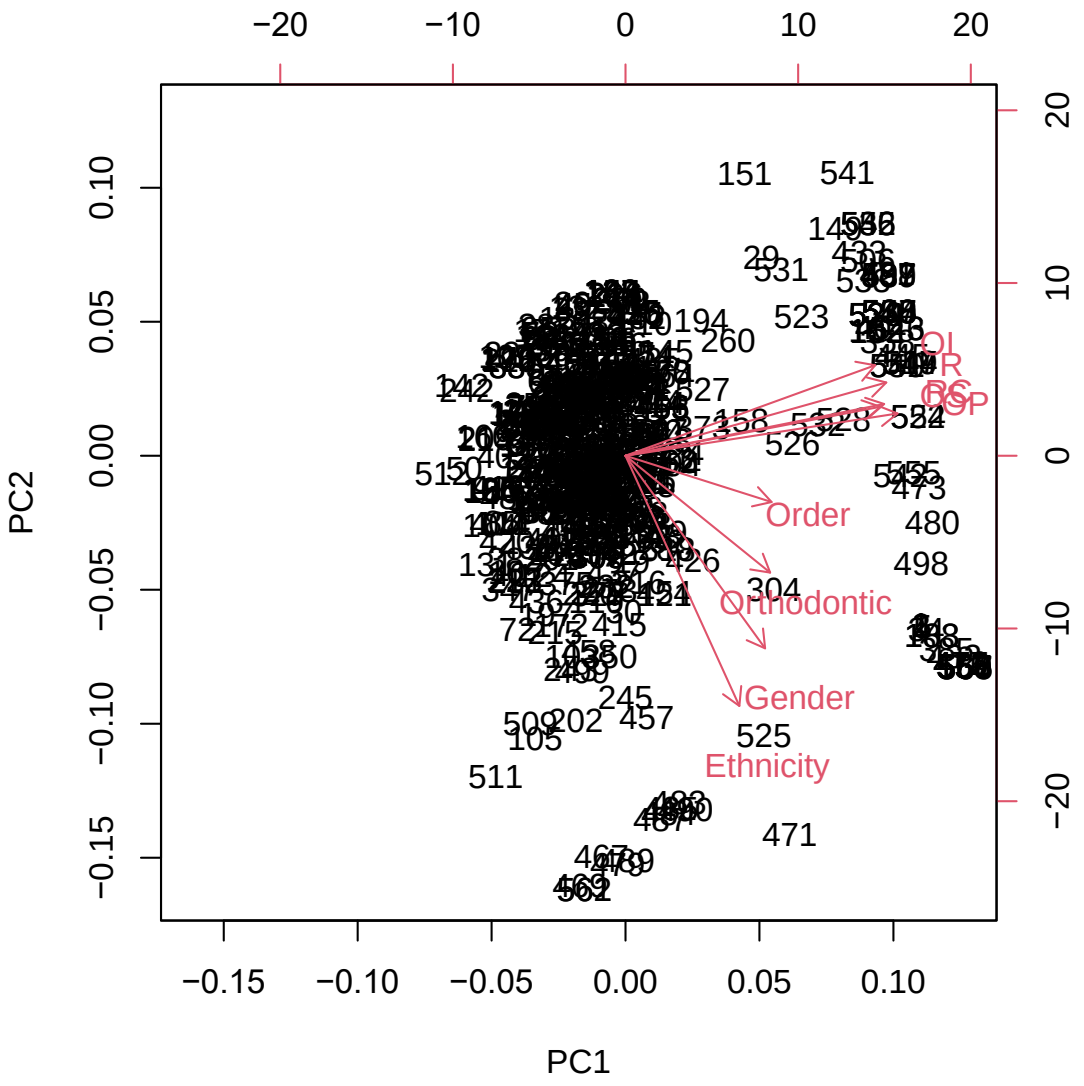

Figure 6. Displays the how order of participated patients was converted to numbers of data to R. The conversion to participant number was necessary in identifying patients in RStudio. By identifying patients in their selected order numbers, their dental and medical histories could be observed.

Figure 7 represents the same corresponding results as previously mentioned in Figure 6, but different scaling of biplot. The different scales of the biplots helped identify participants locations when being analyzed on the biplot. The positions established the parameters of interest

for participants and the parameters which they were closest to in distance on axes. Outliers on right portion of biplot were identified and further observed.

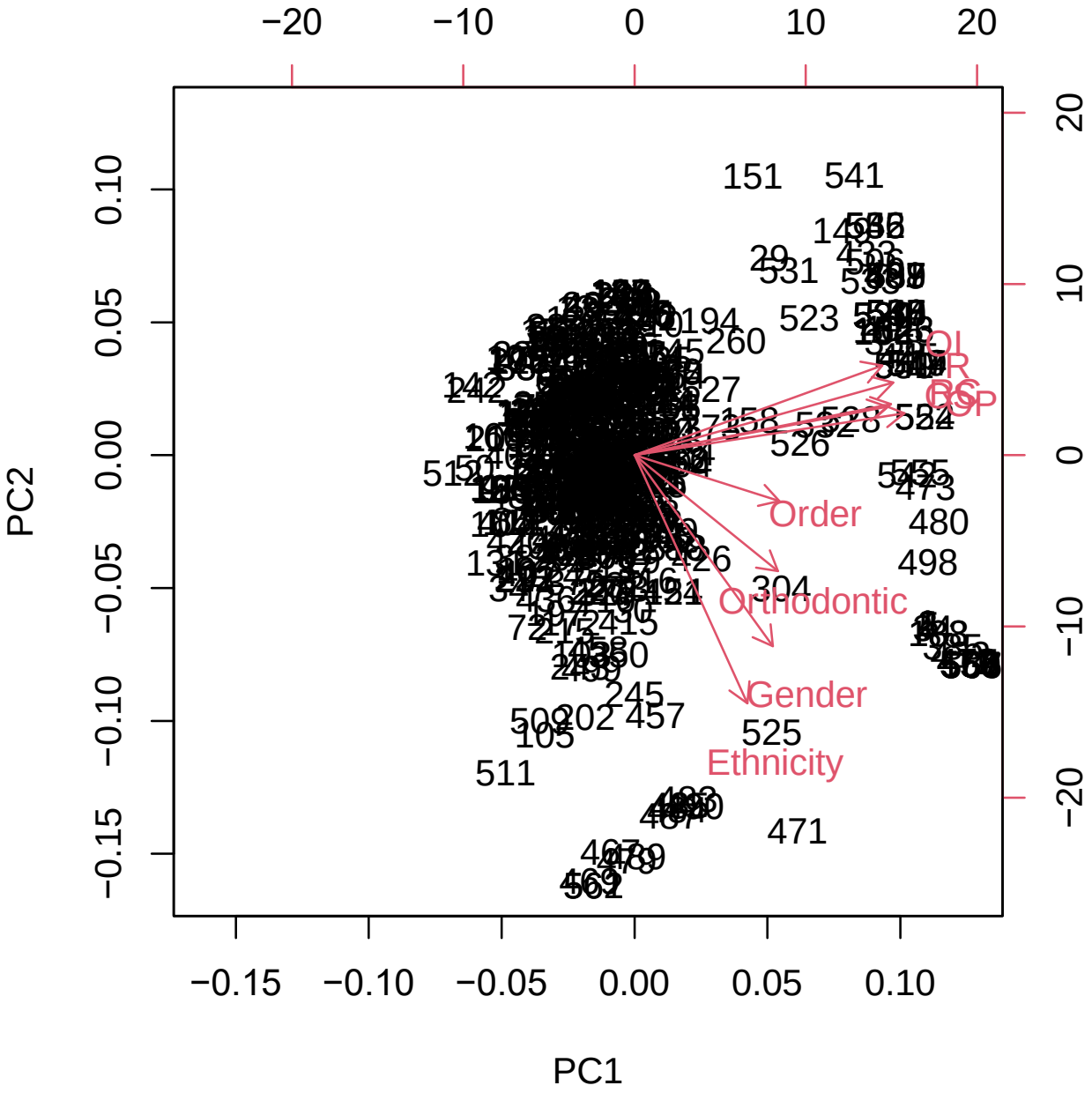

Figure 7. This is a different scaling of biplot of Figure 3. The difference in scaling can identify patients on biplot. The locations allow the variables of interest and participants to be compared. Outliers on right were identified.

After observation, most of the data look similar. This shows that there is not as much diversity to patients as previously predicted. The next step is to gain more parameters to see if a split in data can be established setting pediatric patients apart from each other. PCA was next tested with more parameters including sex, ethnicity, restorations, oral infections, oral surgery, root canals, periodontal, prosthodontics, orthodontics, oral prophylaxis, general health state, has the participant been hospitalized in the last 5 years, is the participant under a physicians care, does the patient take any medications, does the patient have allergies, drug use, alcohol use, tobacco use, history of mental health issue, epilepsy, fainting, stroke, asthma, sinus problems, hepatitis, bruise or bleed easily, irregular heartbeat, high blood pressure, mitral valve prolapse, congenital heart lesions, artificial heart valves, cancer, cancer radiation, heart surgery, kidney disease or dialysis, diabetes, prosthetic joints, and HIV AIDS. Figure 5 displays biplot including the additional parameters being tested with the 528 participants. The split shown in the biplot is further investigated.

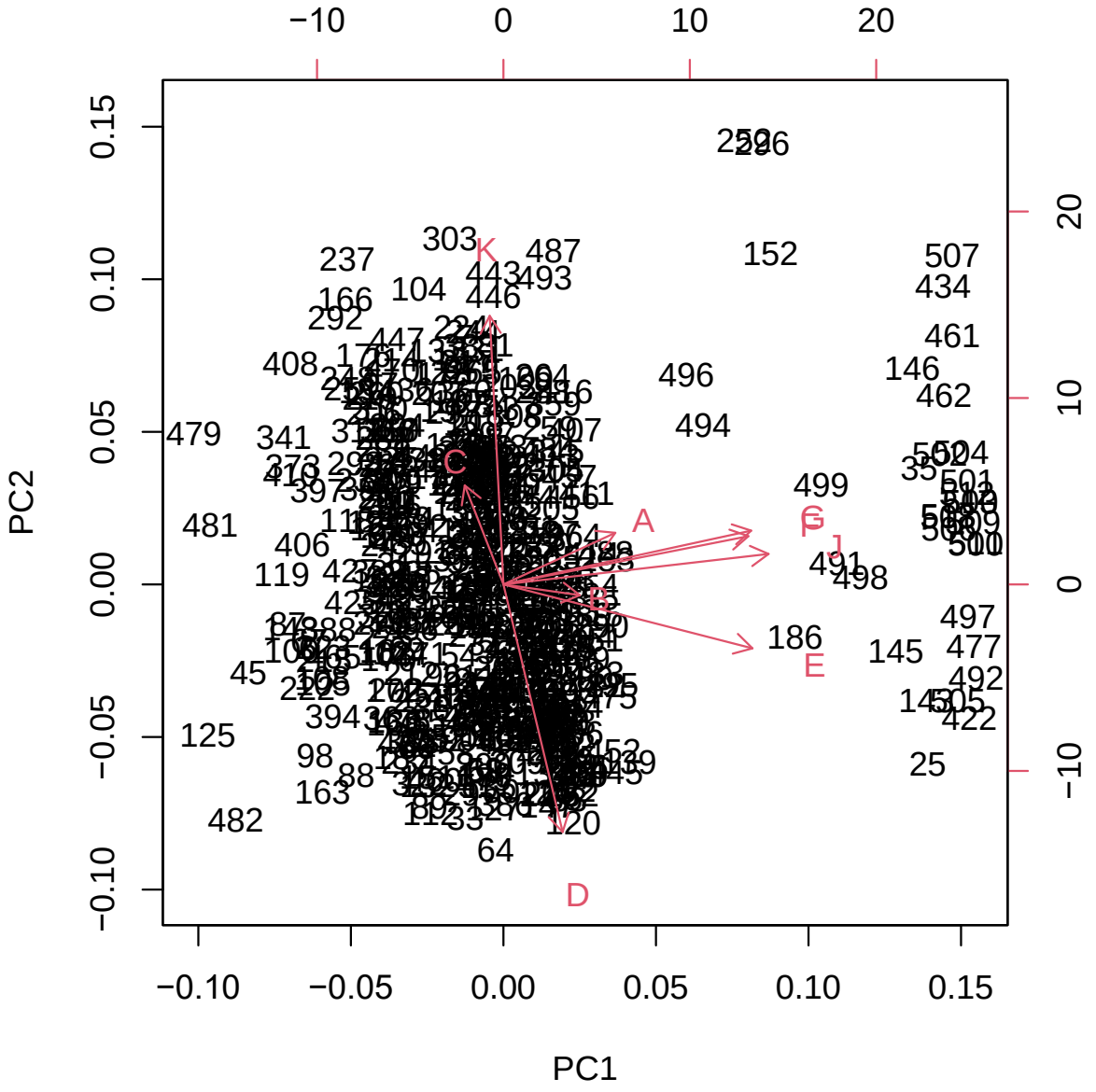

Figure 8. Displays PCA biplot including the additional parameters being tested with the 528 participants. The divide in data shown in the biplot is further observed.

The population is not as diverse as we initially predicted. We thought that there would be more of a drastic difference between the patients with large population sample of 528 patients, but there is only the split in the middle. This correlates with the idea that the children are either not having as medical conditions as predicted and/ or there is not much difference between the medical

conditions and dental conditions of patients or the patients in the pediatric department of the University of Pittsburgh School of Dental Medicine have similar dental issues and medical conditions. This will need to be further investigated to confirm.

The pediatric patients in Figure 8 were then further investigated by observing the split in data. The patients were then identified through RStudio and could be compared to the parameters. The right side of the data that is split will be further explained as the "Northeast side. The "Northeast side" of the divide in Figure 9 biplot displays that most patients' variables that are most similar are sex (B) and restorations (R). The left side of Figure 10 will be further explained as the "Northwest side". The Northwest side patients' variables that are most similar are Oral infections (F) and S (alcohol use). When looking at the comparisons in the Northeast or right side of the divide, sex is predominantly male and shows they had no restorations, but predominantly Oral Infections. When looking at the left side of the split or the Northwest side, sex is predominantly female and shows they had restorations present in their dental histories from the DRDR. This can be shown in the biplots of Figure 9 "Northeast data patients" and Figure 10 "Northwest data patients"

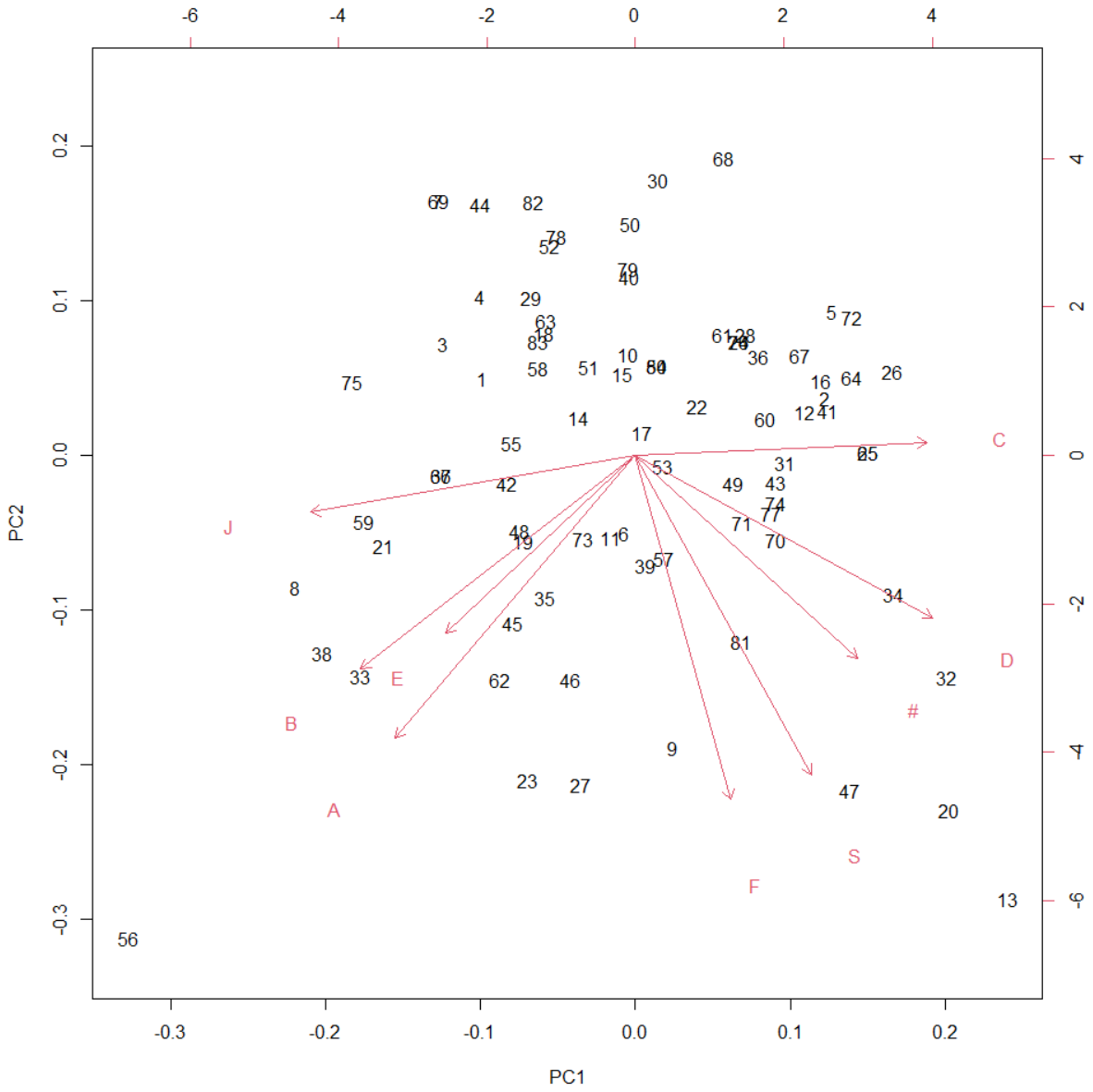

Figure 9. Northeast patients most comparable parameters were sex was predominantly male and showed they had no restorations, but predominantly Oral Infections . Displays that most patients' variables that are most similar are sex (B) and restorations (R).

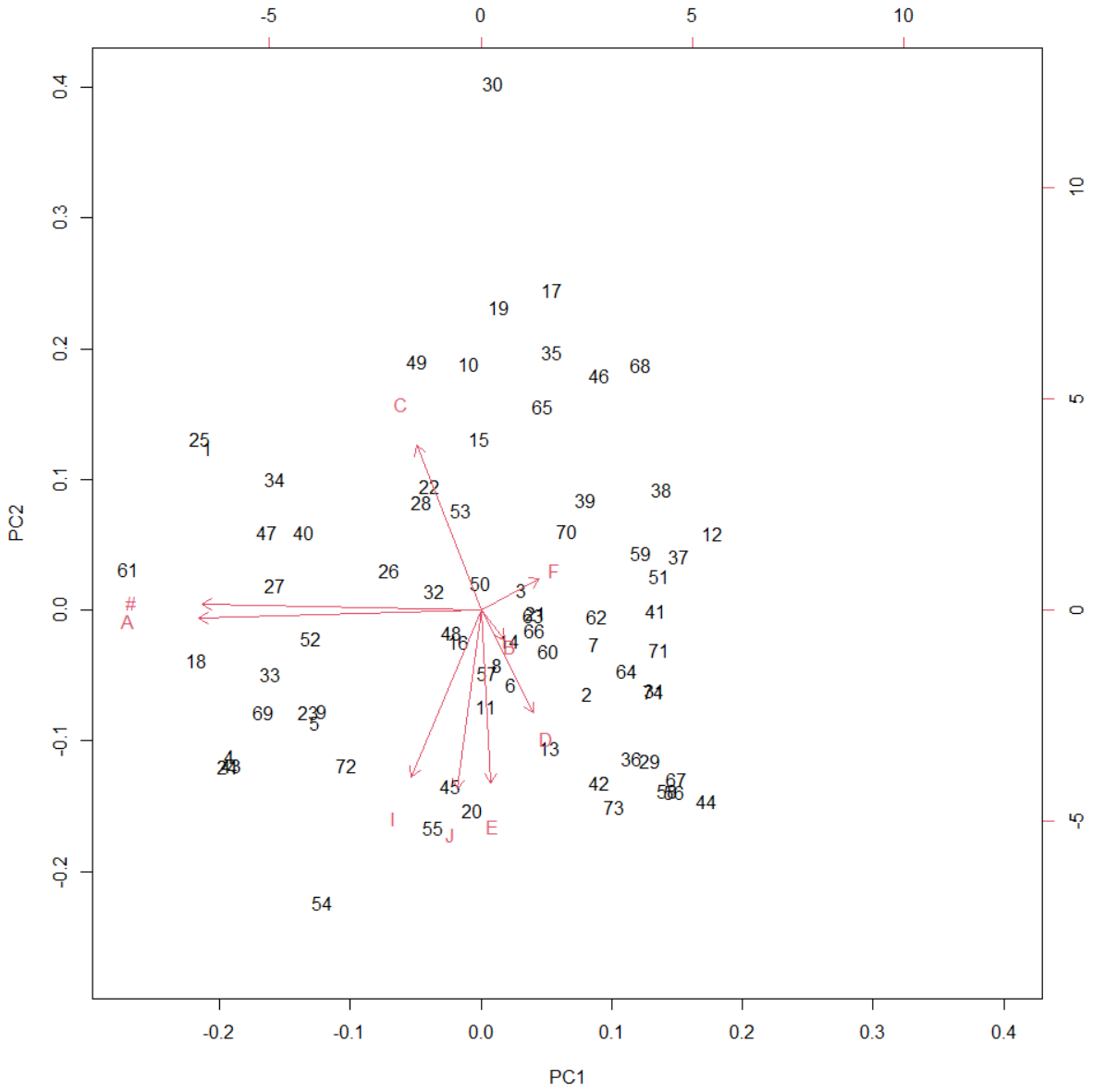

Figure 10. Northwest patients' parameters were sex was predominantly female and showed they had restorations present in their dental histories from the DRDR . The Northwest side patients' variables that were most similar are Oral infections (F) and S (alcohol use).

The patients were separated into groups to test the comparability between the split in the data. The following "Northwest" patients were taken from Figure 10 and tested for four single nucleotide polymorphisms in *MMP2* and *TRAV4*. The following are figures of *TRAV4* and *MMP2*

SNPs in Allelic Discrimination Plot. Figure 14 tested Northwest patients with a fourth marker rs7150049 was tested but was not informative. All patients were homozygous with one exception of a heterozygous (CG).

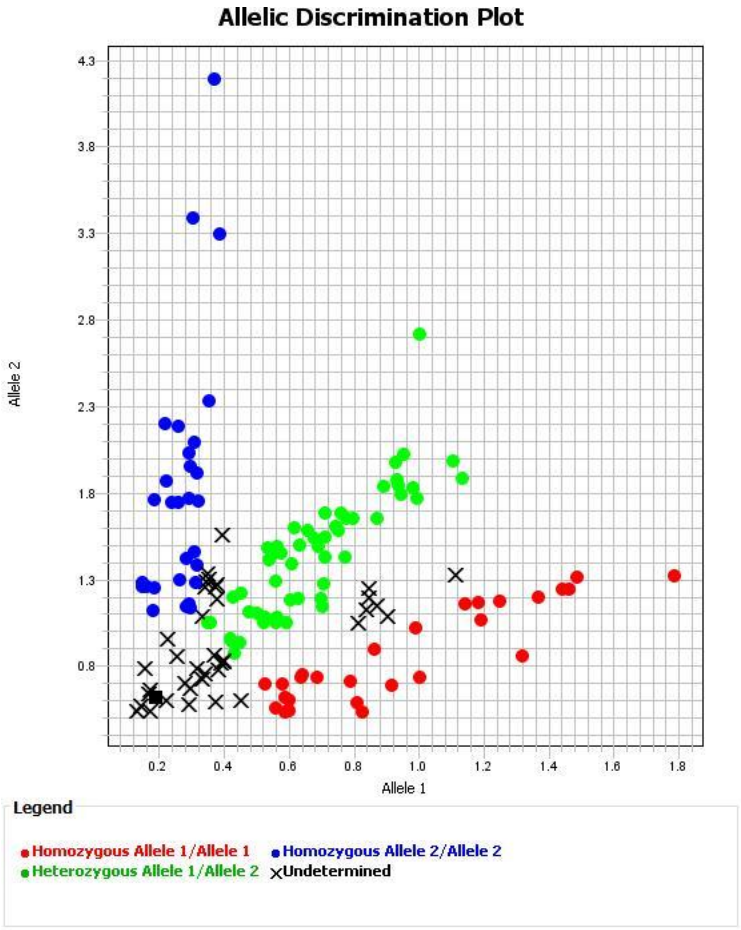

Figure 11. Northwest patients with SNP rs1784418 in *MMP2*

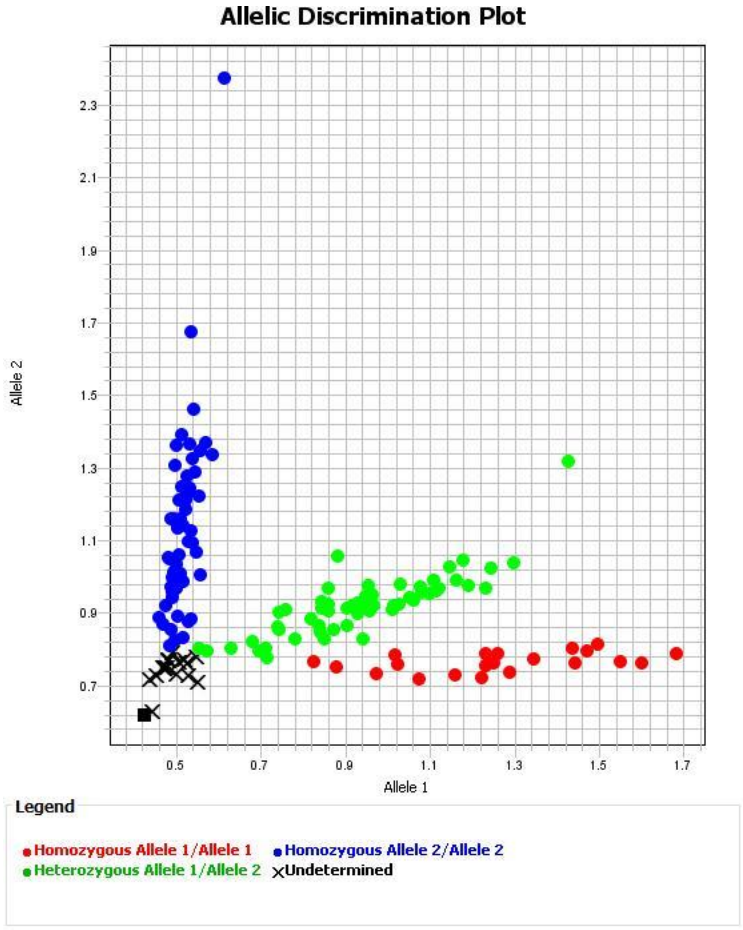

Figure 12. Northwest patients with SNP rs1997533 of *TRAV4*

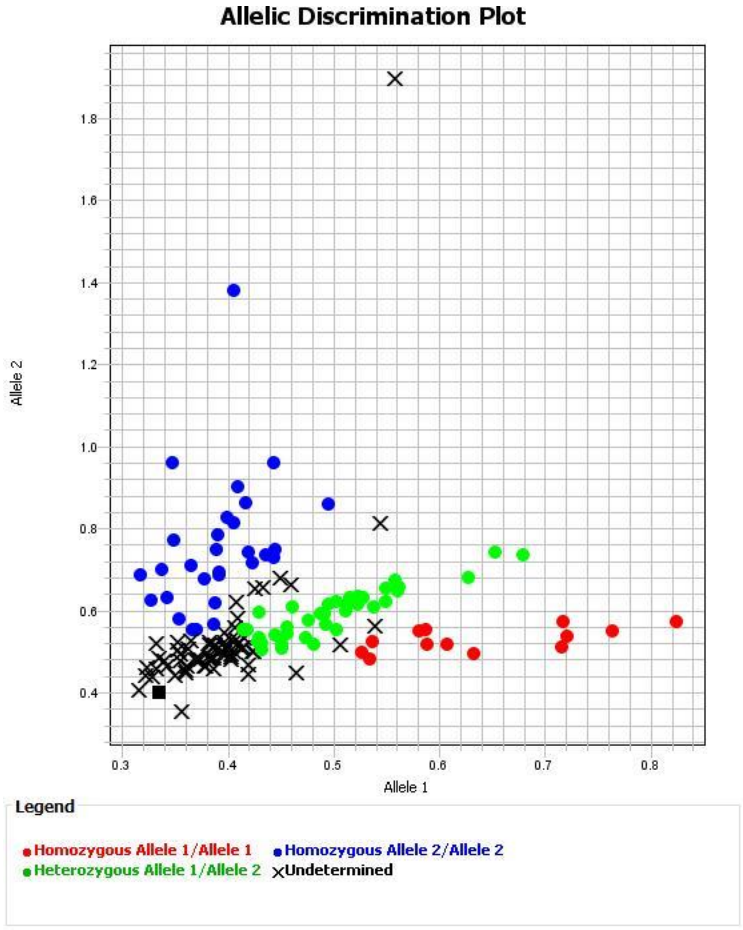

Figure 13. Northwest patients with SNP rs1997533 of *TRAV4*

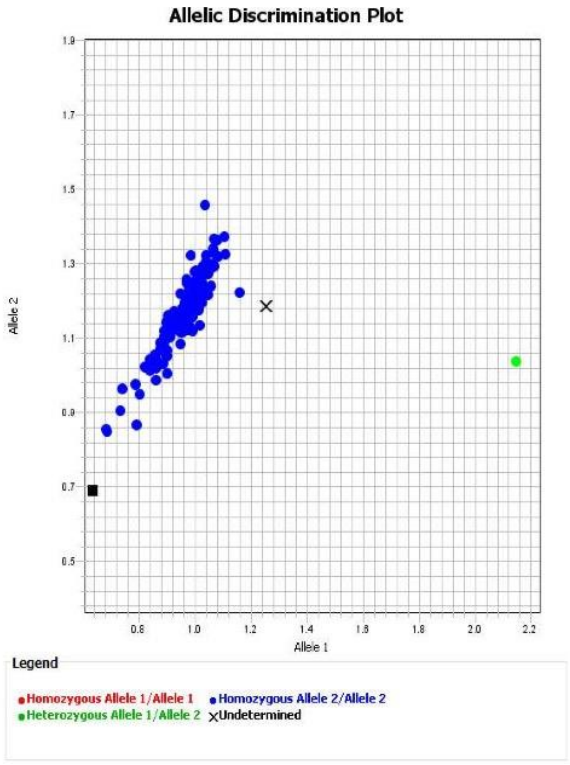

Figure 14. Northwest patients with SNP rs7150049 of *TRAV4*

The following "Northeast" patients were taken from Figure 10 and tested with two single nucleotide polymorphisms, *MMP2* and *TRAV4*.

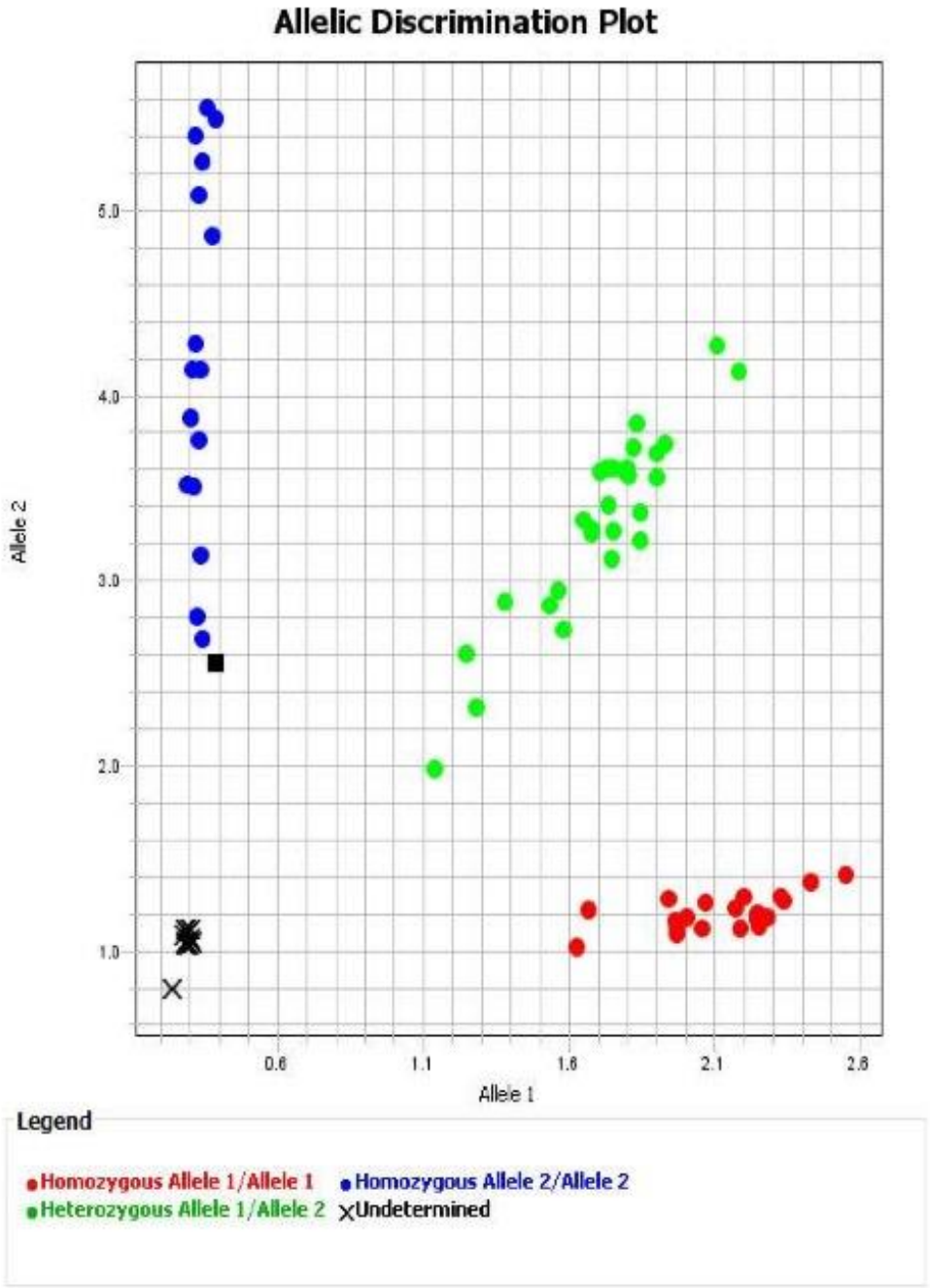

Figure 15. Northeast patients with SNP rs1784418 of *MMP2*.

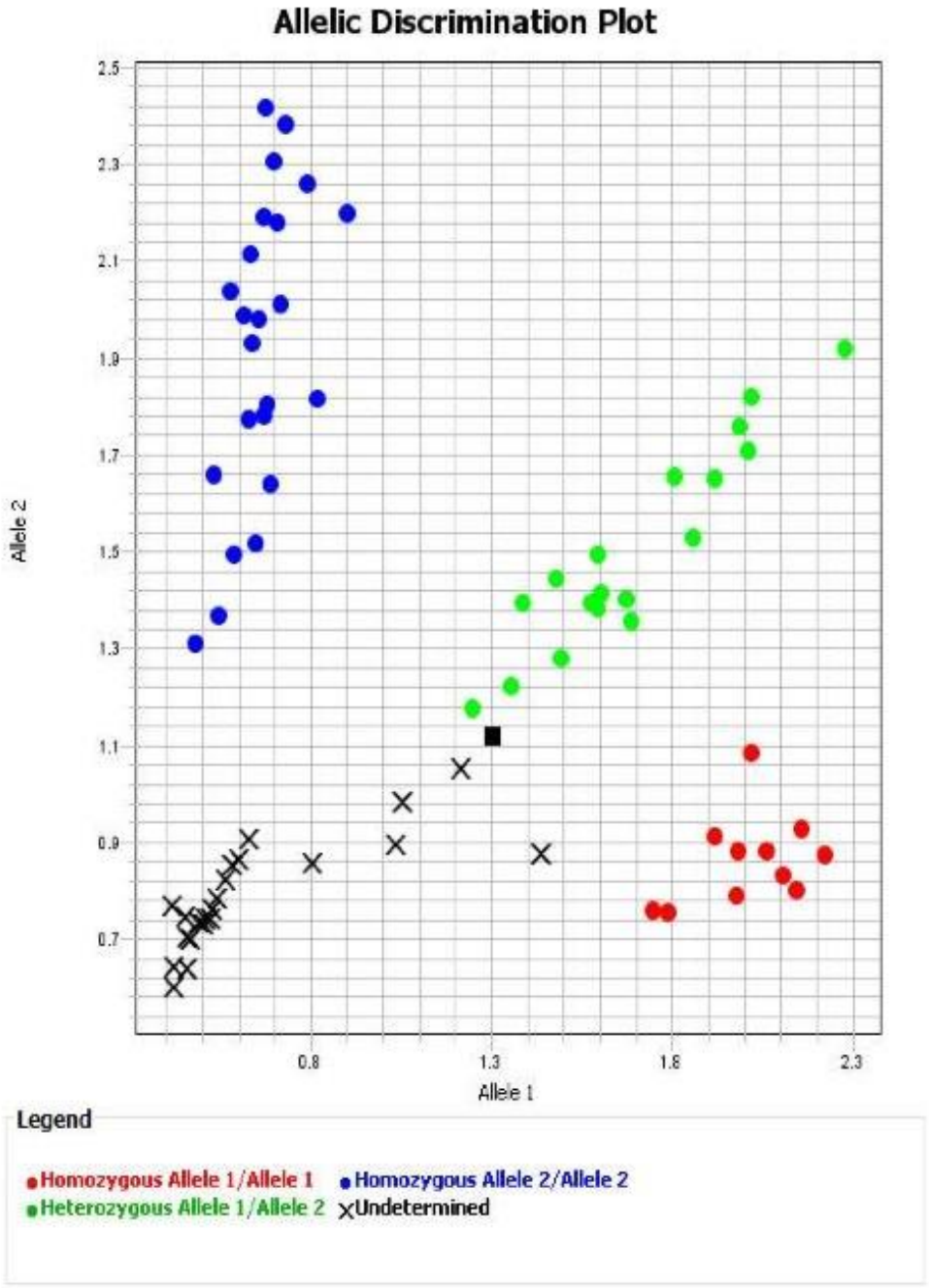

Figure 16. Northeast patients with SNP rs1997532 of *TRAV4*.

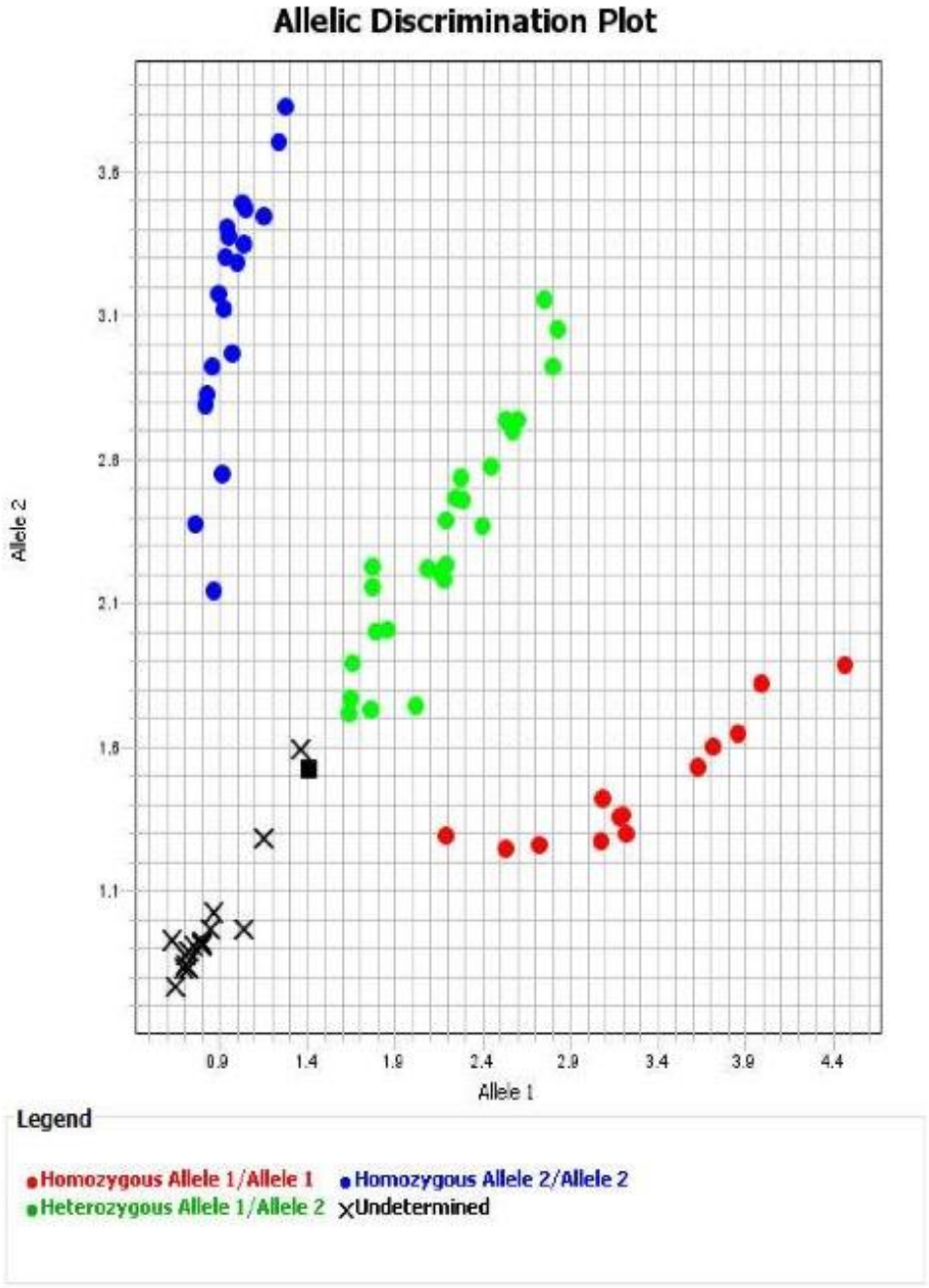

Figure 17. Northeast patients with rs1997533 of *TRAV4*.

The following graphs allowed access to determine patients homozygous and heterozygous alleles. The patients were then separated based on Homozygous Allele 1/Allele 1, Heterozygous Allele 1/Allele 2, and Homozygous Allele 2/ Allele 2.

Table 3. NE and NW patients with genotypes of rs1784418.

rs1784418	CC	CT	TT
NE	22	30	17
NW	28	54	32

Table 4. NE and NW patients with genotypes of rs1997532.

rs1997532	AA	AG	GG
NE	11	17	23
NW	21	57	56

Table 5. NE and NW patients with genotypes of rs1997533.

rs1997533	CC	CG	GG
NE	13	26	19
NW	17	51	36

The following table was for a fourth marker rs7150049 was tested but was not informative. All patients were homozygous with one exception of a heterozygous (CG).

The following graphs were used determine patients homozygous and heterozygous alleles based on their sex. The patients were then separated based on Homozygous Allele 1/Allele 1, Heterozygous Allele 1/Allele 2, and Homozygous Allele 2/ Allele 2 with SNP marker.

Table 6. NW patients with genotypes associated with sex of rs1997532.

Northwest rs1997532	AA	AG	GG
M	17	37	33
F	7	21	19

Table 7. NW patients with genotypes associated with sex of rs1784418.

Northwest rs1784418	CC	CT	TT
M	16	41	14
F	11	13	16

Table 8. NW patients with genotypes associated with sex of rs1997533.

Northwest rs1997533	CC	CG	GG
M	7	24	18
F	8	13	10

Table 9. NE patients with genotypes with sex of rs1784418

Northeast rs1784418	CC	CT	TT
M	9	16	6
F	11	12	10

Table 10. NE patients with genotypes with sex of rs1997532.

Northeast rs1997532	AA	AG	GG
M	5	8	12
F	6	9	11

Table 11. NE patients with genotypes with sex of rs1997533.

Northeast rs1997533	CC	CG	GG
M	6	13	11
F	7	14	8

The following table was for a fourth marker rs7150049 was tested but was not informative. All patients were homozygous (GG) with the exception of one heterozygous marker (CG).

4.1 Statistical Summary

The following table was made from Figures 11-16 and is a summary of statistics between different groups and alleles.

Table 12. Statistical summary of genotypes with rs1784418.

rs1784418	
HW:	p>0.05
Genotype:	p=0.55
Allele:	p=0.32

Table 13. Statistical summary of genotypes with rs1997532.

rs1997532	
HW:	p>0.01
Genotype:	p=0.44
Allele:	p=0.82

Table 14. Statistical summary of genotypes with rs1997533.

rs1997533	
HW:	p>0.05
Genotype:	p=0.63
Allele:	p=0.49

The following table was made from Figures 11-16 and is a summary of statistics between different groups and alleles and sex.

Table 15. Statistical summary of genotypes and sex with rs1784418.

rs1784418	Male	Female
Genotype:	p=0.77	p= 0.65
Allele:	p=0.65	p= 0.35

Table 16. Statistical summary of genotypes and sex with rs1997432.

rs1997532	Male	Female
Genotype:	p=0.60	p= 0.59
Allele:	p=0.54	p= 0.71

Table 17. Statistical summary of genotypes and sex with rs1997433.

rs1997533	Male	Female
Genotype:	p=0.78	p= 0.72
Allele:	p=0.88	p= 0.87

According to Hardy Weinberg equilibrium, it will show if something is evenly distributed in the population. In equilibrium, the alleles are randomly distributed. P value is greater than 0.05 meaning it is not significant. We can trust the genotyping calls obtained. There were no differences in the distribution of alleles and genotypes between the two groups. *TRAV4* and *MMP2* are not associated in a way we initially thought.

5.0 Discussion

Pediatric patients were chosen in this research project to bring more awareness in their difference when being compared to adults. They have oral and craniofacial features that are changing throughout their young life of development and may not be the same in a difference of a couple of years. There are too many young children that suffer from early childhood caries ("ECC") that could have been prevented, suppressed, or arrested through adoption of sustained daily salutary behaviors (Edelstein 2017). ECC remains highly prevalent today affecting children from different socioeconomic backgrounds. The sooner children can begin to get regular dental checkups, the healthier their oral health will be throughout their lives. Early checkups can help prevent cavities and tooth decay, which can lead to other health issues (University of Washington 2020). The primary dentition is temporary but can affect the development of the adult dentition in harmful ways if not taken care of properly. For example, if a baby tooth develops an abscess, the infection could spread to the developing adult tooth (Sebastian Smiles 2019). This could cause early tooth decay in the adult tooth, causing pain and permanent tooth loss (Sebastian Smiles 2019). Pediatric patients require different types of care and more preventative oral health procedures.

The pediatric patients were chosen to be the group observed in this study to strengthen the field of pediatric research. This study initially was looking at the diversity of the pediatric patients at the University of Pittsburgh School of Dental Medicine. Principal component analysis (PCA) was the chosen method to differentiate the patients and was aimed to cluster similar individuals to generate homogeneous groups for future genetic studies. In this research study, we found that patients were more similar in comparison than initially predicted. After observing the similarity in

the patients, further research was conducted to find a distinguishable in the medical history and dental health conditions that could possibly generate a difference in the sample. There were more variables included in the PCA and created a clear separation in the patients. The main area that separated the populations was oral infections, restorations, and sex. The single nucleotide polymorphisms that were chosen to test allelic discrimination and identify genotypes from the population were *TRAV4* and *MMP2*. After completing the statistical summaries of these variables being tested, there were no differences in the distribution of the SNP markers between the two groups.

After identifying two apparently distinct groups using a PCA strategy, the distribution of genetic variants of four markers in two genes did not show any differences, even when sex was considered. When looking at the differences between the two groups from the pediatric patient population, there was a higher prevalence of oral infections in males and restorations in females. An observation that can be made from the results is the bacterial environment for the pediatric patients could potentially be the same. Oral infections and restorations are both two environments for bacteria to thrive in. Dental infections are most commonly occurring when bacteria invade the pulp and spread to surrounding tissues and this can be due to dental caries, trauma, or dental procedures (Erazo et al. 2021). For example, *Streptococcus mutans* is considered as the primary etiologic agent of dental caries, which is an infectious disease (Erazo et al. 2021). These two dental problems can be related by infection caused in the mouth. They are both destructive in retrospect, so that could be related to why there is not as much variation in the population, like we had initially predicted.

Another initial reason there is not as much diversity between the two groups is that there is the same variability between them. There is the same variability between the two groups on

these 4 markers. The variability allows the how far apart the patient's lie from each other and from the center of a distribution (Bhandari 2022). These markers that were chosen, *TRAV4* rs1997532, rs1997533, and rs7150049 and *MMP2* rs1784418, are a few in comparison to the hundreds of thousands of possibilities. In the future, it may be interesting to choose another marker that can help in comparing oral infections In previous study in Poland, they analyzed the genotypes and allele frequencies in five markers single nucleotide polymorphisms (SNPs) in rs17878486 ($p < 0.0001$) in *AMELX*, rs34538475 in *AMBN* ($p < 0.0001$), rs2337360 in *TUFT1* ($p < 0.0001$), and rs2235091 ($p = 0.0085$) and rs198969 ($p = 0.0069$) in *KLK4* genes that proved they were significantly associated with dental caries occurrence in a population of Polish children (Gerreth et al.). Another study in Finland found SNPs in *DDX39B* and MPO showed association tendencies between caries and the genes from Finnish adolescents (Raivisto 2018). These similar studies represent other of the hundreds of markers that can be further tested in the pediatric patient population at the University of Pittsburgh School of Dental Medicine.

The next step in further investigating the pediatric patient's differences would be observing sex. Single nucleotide polymorphisms could be specifically chosen based on the patient's sex and investigated. This could be potentially what could create variability in the population. Sex could be a distinguishing factor in the medical history and dental health conditions. Previous studies have looked at genetic and environment factors that may have caused dental caries and oral infections in the mouth, but no studies have looked at the sex of the pediatric patient and dental caries and oral infections. This could be an interesting next step in determining why the two groups may differ.

There have been differences acknowledged in literature looking at caries of patients based on their sex. Many studies have demonstrated that caries rates are higher in women than in men

(Ferraro and Vieira 2009). Evidence has been provided to demonstrate that caries risk factors for women include a different salivary composition and flow rate, hormonal fluctuations, dietary habits, genetic variations, and particular social roles among their families (Ferraro and Vieira 2009). This could also be due to the idea that females may report discomfort or issues in their mouth before males due. Females may feel a certain pressure of appearance that males may not be so inclined, especially in the age group that is being compared. These findings correlate with the results of the research study, where females had majority of restorations from the PCA results.

We cannot further explain why the two groups may differ, which would explain the PCA initial findings. The population may be more similar than we thought. This study brings attention to the necessity for more research to be done in the field of pediatric dentistry. There is such a difference in pediatric dental research versus adult dental research. From this study, we can say the pediatric patient population at the University of Pittsburgh School of Dental Medicine is not as diverse from statistical testing. This is important because possibly children may not have as much diversity in the Appalachia region as one would initially predict.

It was important to understand variables and factors that could correlate with their dental health conditions and treatments for the pediatric subjects chosen. In this research experiment, we selected different parameters from pediatric patients medical and dental health evaluations to see if there is any similarity in specific medical factors influencing dental health conditions. This was the first aim to establish a sample population to compare to one another. The data of the initial 528 patients included demographics, self-reported medical history, medications taken, and oral health conditions that were reported by the pediatric subjects. The population size of the patients provided accurate mean values, the identity of outliers that could possibly skew the data versus using a

smaller sample and provide a smaller margin of error (Zamboni 2018). The larger samples more closely approximate the population (Jason 2007).

Principal component analysis (PCA) was used to cluster similar individuals aiming to generate homogeneous groups for future genetic studies. This was the goal of second aim to use the medical factors and dental health conditions noted by the pediatric subjects as variables. PCA is a excellent way to establish a comparable figure to illustrate similarities and differences based on data from a sample population. The pediatric subjects were able to be compared to each other by their medical history and dental health conditions. This allowed the identity of the patients for further investigation of the separation in the graph.

It is important to note that patients were able to answer N/A for specific medical and dental questions to be listed. N/A can have multiple different meanings such as Not Applicable, Not Accessible, Not Appropriate, etc. This could correlate to potentially missing data or information that could separate the population from one another. The variables were assigned a number "3" to establish N/A answer. It would be interesting to eliminate the patients that have the "3" in their medical and dental records and see what separates them from one another.

One of the predominant reasons that could be that influencing the similarities in the pediatric patients is the demographic of the region. It is important to understand the profile of the children in the region being tested in this study. The profile of children in this study is from Pittsburgh, USA. Pittsburgh is the largest city in the Appalachian region which is one of the poorest areas in the country (Vieira et al. 2020). This area is known to have some of the worse health outcomes of the country (Vieira et al. 2020). The city of Pittsburgh has lots of different areas within it that have different percentages that are below federal poverty line and are below the census tract. After noting the demographic of the region and looking at the PCA graph, we can understand why

majority of the procedure arrows are close in proximity. The pediatric patients are more similar in their medical and dental conditions due to their current profile and demographic which they live in. This makes sense why we did not see as much of a split in the population when comparing their medical and dental health records.

We found that patients were more similar in comparison than initially predicted. This difference can be further observed by distinguish the medical factors that caused a difference in the final group sample dental health conditions. This was the goal of Aim 3 to investigate the split in the population by using single nucleotide polymorphisms *TRAV4* and *MMP2* when comparing the separations in populations. We were able to identify genotypes of pediatric patients when using Real Time Allelic Discrimination with SNP markers rs1997532, rs1997533, rs1784418, and rs7150049.

We were able to take the genotypes that were found from the pediatric patients Real Time Allelic Discrimination with SNP markers rs1997532, rs1997533, rs1784418, and rs7150049 for statistical testing. The chi-square was used to test for Hardy-Weinberg equilibrium. The p-values below 0.001 were not in equilibrium. Chi-square was also used for testing for over-representation of genotypes or alleles and sex between the two groups with an alpha of 0.05. The goal was to determine if the difference in the distribution of the alleles between the two groups could be attributed to influences from *TRAV4* or *MMP2* on susceptibility to oral conditions. There were no differences in the distribution of alleles and genotypes between the two groups.

This study does have some limitations. The limitations are the little variability of the data but establishes similarity between the patient's variables being tested through PCA. Also, variables such as oral hygiene are not very informative because all kids may also receive it, no matter the evaluation or diagnosis of their oral health. The sample size of the population could always be

larger to have a larger amount of variation, but this is the amount of patient's available to compare from the DRDR. And finally, the registry design always implicates some inconsistency on how data can be recorded. We could not control these limitations but are important to note to the study.

6.0 Conclusion

1 528 pediatric subjects 0-18 years of age were chosen to observe if there is any similarity in specific medical factors influencing dental health conditions between the different parameters from pediatric patients medical and dental health evaluations. There is no further evidence that explains why the two groups may differ, which would explain the PCA initial findings. The population is more similar than we thought. After, we can say the pediatric patient population at the University of Pittsburgh School of Dental Medicine is not as diverse from statistical testing.

2 Principal Component Analysis did allow for a clear separation in pediatric subjects, but it was unclear what markers caused this divide. It also provided the representation of the identity of groups of pediatric patients. This allowed us to find which variables makes the pediatric subjects different from another and which variables are similar. The intended goals of PCA were achieved in extracting the most important variables from the data table, which was comparing sex and oral infections and restorations, allowing the compression of the size and simplifying the data set of the 528 pediatric patients into two identifiable groups to further analyze, and analyzing the overall structure from the PCA biplot of the observations and the parameters.

3 The distribution of alleles and genotypes of variants in *TRAV4* and *MMP2* were not statistically significant to when comparing the pediatric subjects. These two genes *TRAV4* or *MMP2* where chosen based on susceptibility to oral infection stated in medical history and dental records. They were determining if the difference in the distribution of the alleles between the two groups could be attributed to the parameters found in the separation of the population in the Principal Component Analysis. Ultimately, there were no differences in the distribution of alleles and genotypes between the two groups.

Bibliography

Abbasoğlu, Z., Tanboğa, İ., Küchler, E. C., Deeley, K., Weber, M., Kaspar, C., Korachi, M., & Vieira, A. R. (2015). Early childhood caries is associated with genetic variants in enamel formation and immune response genes. *Caries research*, *49*(1), 70–77. https://doi.org/10.1159/000362825

Abdi, H. and Williams, L.J. (2010), Principal component analysis. WIREs Comp Stat, 2: 433-459. https://doi.org/10.1002/wics.101

Antunes, L. A., Antunes, L. S., Küchler, E. C., Lopes, L. B., Moura, A., Bigonha, R. S., Abreu, F. V., Granjeiro, J. M., de Amorim, L. M., & Paixão, I. C. (2015). Analysis of the association between polymorphisms in*mmp2*,*mmp3*,*mmp9*,*mmp20*,*timp1*, and*timp2*genes with white spot lesions and early childhood caries. *International Journal of Paediatric Dentistry*, *26*(4), 310–319. https://doi.org/10.1111/ipd.12202

Benli, M., Frota de Souza, L. A., Deeley, K., Modesto, A., & Vieira, A. R. (2021). Matrix metalloproteinase 2 is associated with secondary caries independent from the restorative material. *Frontiers in Dental Medicine*, *2*. https://doi.org/10.3389/fdmed.2021.735535

Bhandari, P. (2022, January 31). *Measures of variability*. Scribbr. Retrieved from https://www.scribbr.com/statistics/variability/#:~:text=Variability%20tells%20you%20how%20far,as%20spread%2C%20scatter%20or%20dispersion.

Braveman P, Egerter S, Williams DR. The social determinants of health: coming of age. *Annu Rev Public Health*. 2011;32:381– doi:10.1146/annurev-publhealth-031210-101218

Briseño-Ruiz, J., Shimizu, T., Deeley, K., Dizak, P. M., Ruff, T. D., Faraco, I. M., Jr, Poletta, F. A., Brancher, J. A., Pecharki, G. D., Küchler, E. C., Tannure, P. N., Lips, A., Vieira, T. C., Patir, A., Koruyucu, M., Mereb, J. C., Resick, J. M., Brandon, C. A., Letra, A., Silva, R. M., … Vieira, A. R. (2013). Role of TRAV locus in low caries experience. *Human genetics*, *132*(9), 1015–1025. https://doi.org/10.1007/s00439-013-1313-4

Centers for Disease Control and Prevention. (2021, February 5). *Disparities in oral health*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/oralhealth/oral_health_disparities/index.htm

Edelstein B. L. (2017). Pediatric Dental-Focused Interprofessional Interventions: Rethinking Early Childhood Oral Health Management. *Dental clinics of North America*, *61*(3), 589–606. https://doi.org/10.1016/j.cden.2017.02.005

Erazo D, Whetstone DR. Dental Infections. [Updated 2021 Oct 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK542165/

Fanchon, S., Bourd, K., Septier, D., Everts, V., Beertsen, W., Menashi, S. and Goldberg, M. (2004), Involvement of matrix metalloproteinases in the onset of dentin mineralization. European Journal of Oral Sciences, 112: 171-176. https://doi.org/10.1111/j.1600-0722.2004.00120.x

Ferraro, M., & Vieira, A. R. (2010). Explaining gender differences in caries: A multifactorial approach to a multifactorial disease. *International Journal of Dentistry*, *2009*, 1–5. https://doi.org/10.1155/2010/649643

Fisher-Owens SA, Gansky SA, Platt LJ, Weintraub JA, Soobader MJ, Bramlett MD, Newacheck PW. Influences on children's oral health: a conceptual model. *Pediatrics.* 2007;120:e510–520.

freeCodeCamp.org. (2019, April 6). *An overview of principal component analysis*. freeCodeCamp.org. Retrieved from https://www.freecodecamp.org/news/an-overview-of-principal-component-analysis-6340e3bc4073/

Gerreth, K., Zaorska, K., Zabel, M., Borysewicz-Lewicka, M., & Nowicki, M. (2017). Chosen single nucleotide polymorphisms (SNPs) of enamel formation genes and dental caries in a population of Polish children. *Advances in clinical and experimental medicine : official organ Wroclaw Medical University*, *26*(6), 899–905. https://doi.org/10.17219/acem/63024

Ggbiplot. RDocumentation. (n.d.). Retrieved from https://www.rdocumentation.org/packages/ggbiplot/versions/0.55

Hayden, L. (2018, August 9). *R PCA tutorial (principal component analysis)*. DataCamp Community. Retrieved from https://www.datacamp.com/community/tutorials/pca-analysis-r

Holand, S. M. (2019, December 3). *Principal components analysis PCA - University of Georgia*. University of Georgia, Athens, GA 30602-2501. Retrieved from https://strata.uga.edu/software/pdf/pcaTutorial.pdf

Jason. (2007). *Sample Size*. Lecture4. Retrieved from http://web.pdx.edu/~newsomj/pa551/lecture4.htm

Jolliffe Ian T. and Cadima Jorge, Ian T. Jolliffe Ian T. Jolliffe College of Engineering, Jolliffe, I. T., Ian T. Jolliffe College of Engineering, Cadima, J., Jorge Cadima Secção de Matemática (DCEB), & One contribution of 13 to a theme issue 'Adaptive data analysis: theory and applications'. (2016, April 13). *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

Lima, F. (2017, January 23). *Principal component analysis in R: R-bloggers*. R. Retrieved from https://www.r-bloggers.com/2017/01/principal-component-analysis-in-r/

Menezes-Silva, R., Khaliq, S., Deeley, K., Letra, A., & Vieira, A. R. (2012). Genetic susceptibility to periapical disease: conditional contribution of MMP2 and MMP3 genes to the

development of periapical lesions and healing response. *Journal of endodontics*, *38*(5), 604–607. https://doi.org/10.1016/j.joen.2012.02.009

National Institute of Dental and Craniofacial Research. *Oral Health in America: A Report of the Surgeon General*. National Institutes of Health, US Department of Health and Human Services; 2000.

Niu LN, Zhang L, Jiao K, Li F, Ding YX, Wang DY, et al. Localization of MMP-2, MMP-9, TIMP-1, and TIMP-2 in human coronal dentine. *J Dent.* (2011) 39:536–42. doi: 10.1016/j.jdent.2011.05.004

PCA Analysis in R. DataCamp Community. (n.d.). https://www.datacamp.com/community/tutorials/pca-analysis-r?utm_source=adwords_ppc&utm_campaignid=1565261270&utm_adgroupid=67750485268&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332661264374&utm_targetid=aud-392016246653%3Adsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9005925&gclid=Cj0KCQjw-O35BRDVARIsAJU5mQVZcd8xaRvGMQNmpM6G19bx_jgmE3eKbS-fdP846qazuYBHaD6_PeAaAk7TEALw_wcB

Powell LV. Caries prediction: a review of the literature. *Community Dent Oral Epidemiol.* 1998;26:361–371

Raivisto, T., Heikkinen, A., Kovanen, L., Ruokonen, H., Kettunen, K., Tervahartiala, T., Haukka, J., & Sorsa, T. (2018). SNP Analysis of Caries and Initial Caries in Finnish Adolescents. *International journal of dentistry*, *2018*, 1586762. https://doi.org/10.1155/2018/1586762

REDCap. (n.d.). https://projectredcap.org/about/.

Smiles, S. (2019, July 1). *Why baby teeth matter for an adult smile*. Sebastian Pediatric. Retrieved from https://sebastiansmilespediatricdentistry.com/2019/07/01/why-baby-teeth-matter-for-an-adult-smile/#:~:text=Untreated%20Primary%20Tooth%20Decay%20May%20Affect%20Adult%20Teeth&text=If%20a%20baby%20tooth%20develops,of%20your%20child's%20primary%20teeth.

Tannure, P. N., Küchler, E. C., Lips, A., Costa, M., Luiz, R. R., Granjeiro, J. M., & Vieira, A. R. (2012). Genetic variation in MMP20 contributes to higher caries experience. *Journal of dentistry*, *40*(5), 381–386. https://doi.org/10.1016/j.jdent.2012.01.015

Thiagogm. (2013, November 28). *Computing and visualizing PCA in R: R-bloggers*. R. Retrieved from https://www.r-bloggers.com/2013/11/computing-and-visualizing-pca-in-r/

Vieira A, R, Modesto A, Marazita M, L: Caries: Review of Human Genetics Research. Caries Res 2014;48:491-506. doi: 10.1159/000358333

Vieira, A. R. (2020). The future research agenda of pediatric dentistry. *Frontiers in Dental Medicine*, *1*. https://doi.org/10.3389/fdmed.2020.00004

Wang, X., Willing, M. C., Marazita, M. L., Wendell, S., Warren, J. J., Broffitt, B., Smith, B., Busch, T., Lidral, A. C., & Levy, S. M. (2012). Genetic and environmental factors associated with dental caries in children: the Iowa Fluoride Study. *Caries research*, *46*(3), 177–184. https://doi.org/10.1159/000337282

Why early childhood dental visits are important. The Center for Pediatric Dentistry. (2020, October 8). Retrieved from https://thecenterforpediatricdentistry.com/for-parents-and-patients/why-early-childhood/

World Health Organization. *The world oral health report 2003.* 2003.

Wright, T. J. (2020, April 14). *The role of genetics in caries risk and resistance*. The Role of Genetics in Caries Risk and Resistance. Retrieved from https://dimensionsofdentalhygiene.com/article/genetics-caries-risk/

Zamboni, J. (2019, March 2). *The advantages of a large sample size*. Sciencing. Retrieved from https://sciencing.com/advantages-large-sample-size-7210190.html