Investigation of a gel effect on dental disease by Generalized Linearized Model

Introduction:

In this project, the investigators were interested in a new gel treatment for gum disease on 130 participants. The hypothesis of interest was whether the gel treatment would lower the pocket depth or attachment loss at one year. There were 4 categorical variables, treatment group, race, sex and smoker; and 6 numerical variables, age, sites, attachbase, attach1year, pdbase and pd1year in the dataset. Since there are very high correlations between "the baseline score" and "the score at one year", hence we generated two new variables, "attachchange" and "dpchange", in order to measure the changes of the pocket depth and attachment loss adjusting other factors. The diagnostic plots showed the data has missing data, but it is hard to impute the accurate data since there was very limited data points. GLM and T-test were applied for comparison to test the treatment significance among the different groups and between two different groups. GLMselect by stepwise selection was used to generate the regression linear model, which confirmed the result of GLM, and indicated the treatments were not significant.

Methods:

Since there were missing data which were noted by "NA" in the dataset, the data were difficult to properly be imported in the SAS Studio (although it can be easily read and changed in R Studio). I had to change the NA into "." in the Excel file before reading the data into SAS Studio.

The study interest was **whether treatment results in lower average pocket depth and attachment loss at one year.** The score difference/change was the interested

outcome for this research. Hence two new variables, attachchange and pdchange, were generated for data cleaning. In order to select the variables in the model, the data was investigated for correlations between the variables by “Proc Corr”.

Then the missing data pattern was showed that age had one missing data. Some of the participants were difficult to be followed up after 1 year by various reasons. There were 27 data points missing in attach loss at one year and pocket depth at one year.

The Gplot was to check the relationship between the outcomes and the variables. The numerical variables including age, sites, attachbase, attach1year, pdbase and pd1year are normally distributed or spread evenly, so we don’t have to transform the data.

The mean and frequency were checked separately for the numerical variables and categorical variables in order to get the basic statistics for the data. The treatment groups were random selected and balanced, and there are 26 participants in each treatment group; sex and smoker were balanced too. There is 87% white in the race, although highly weighted than other three populations, randomly balanced dispersed in the treatment groups.

Since there were categorical variables and numerical variables in the dataset, the Generalized Linear Model (GLM) was used to check if the different treatment groups had the different effects on the two outcomes, pdchange and attachchange without the adjustment at the beginning. Fourthly, the Structured Query Language (SQL) was used to generated into different groups in pair and t-test was applied to test which two treatment groups could have different effects for two outcomes.

GLMSELECT was used to select the models and detect the potential significant interactions among the variables. The final regression models were obtained by using the selected variables from the GLMSELECT results. This step with interaction might complicate the study since the sample size was small in each treatment group. Without interaction, the final model was selected and treatment groups were excluded.

Results:

Correlation was check in the Form 1. There were very high correlations between the variables, attachbase and attach1year (r=0.946), pdbase and pd1year(r=0.843). But there were relative low correlations between the variables, attachchange and attachbase (r=-0.414), also pdchange and pdbase (r=-0.203). Hence the two variables, attachchange and pdchange, were selected as the new outcomes for the models.

From 1: Correlations between all the variables

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	pdchange	attachchange	age	sites	pdbase	pd1year	attachbase	attach1year
pdchange	1.00000 103	0.53546 <.0001 103	-0.07422 0.4585 102	-0.04812 0.6293 103	-0.20267 0.0401 103	0.35543 0.0002 103	-0.03676 0.7124 103	0.15128 0.1272 103
attachchange	0.53546 <.0001 103	1.00000 103	-0.17458 0.0793 102	0.16209 0.1019 103	-0.13472 0.1749 103	0.16531 0.0952 103	-0.41437 <.0001 103	-0.09561 0.3367 103
age	-0.07422 0.4585 102	-0.17458 0.0793 102	1.00000 129	-0.11266 0.2037 129	-0.11097 0.2106 129	-0.12526 0.2097 102	0.12162 0.1698 129	0.08605 0.3898 102
sites	-0.04812 0.6293 103	0.16209 0.1019 103	-0.11266 0.2037 129	1.00000 130	-0.16634 0.0586 130	-0.18852 0.0565 103	-0.39708 <.0001 130	-0.36885 0.0001 103
pdbase	-0.20267 0.0401 103	-0.13472 0.1749 103	-0.11097 0.2106 129	-0.16634 0.0586 130	1.00000 130	0.84327 <.0001 103	0.58869 <.0001 130	0.60498 <.0001 103
pd1year	0.35543 0.0002 103	0.16531 0.0952 103	-0.12526 0.2097 102	-0.18852 0.0565 103	0.84327 <.0001 103	1.00000 103	0.54983 <.0001 103	0.66049 <.0001 103
attachbase	-0.03676 0.7124 103	-0.41437 <.0001 103	0.12162 0.1698 129	-0.39708 <.0001 130	0.58869 <.0001 130	0.54983 <.0001 103	1.00000 130	0.94556 <.0001 103
attach1year	0.15128 0.1272 103	-0.09561 0.3367 103	0.08605 0.3898 102	-0.36885 0.0001 103	0.60498 <.0001 103	0.66049 <.0001 103	0.94556 <.0001 103	1.00000 103

Missing data analysis was shown in Form 2. There was one missing (0.77%) in age and 27 missing (20.77%) in pd1year and attach1year, which caused the same missing in the pdchange and attachchange. The reason for the missing was difficult for following up at the one year when the study ended. Since the percentage of the missing data was not

high and could be tolerate, and there were only two time points which was difficult to predict the trend at the one year, we analyzed the data without imputation for the missing data.

Form 2: Missing data Descriptive for Continuous Variables

The MEANS Procedure

Variable	N	N Miss
pdchange	103	27
attachchange	103	27
age	129	1
sites	130	0
pdbase	130	0
pd1year	103	27
attachbase	130	0
attach1year	103	27

The MI Procedure

Missing Data Patterns

Group	pdchange	attachchange	age	sites	pdbase	pd1year	attachbase	attach1year	Freq	Percent
1	X	X	X	X	X	X	X	X	102	78.46
2	X	X	.	X	X	X	X	X	1	0.77
3	.	.	X	X	X	.	X	.	27	20.77

Overall statistical description for variables was shown in Form 3-1 and 3-2. The two outcomes, pdchange and attachchange, had small mean values and small errors (standard deviation). The ages of the participants were ranged from 28 years old to 74 years old. The checked sites had broader range and collected enough numbers for statistical analysis. From 3-2 showed the smoker, age, treatment groups were well balanced, however, 87% white was in the race, and other three populations only had 13% total. The small sample number 26 in each treatment group indicated the study population was small.

Form 3-1: The Means of the continuous variables

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
pdchange	103	-0.2943543	0.2676105	-0.8580247	0.4551282
attachchange	103	-0.0994510	0.2760304	-1.0476190	0.4523810
age	129	49.9434708	10.0323275	28.5722108	74.5325120
sites	130	157.5076923	11.3412509	114.0000000	168.0000000
pdbase	130	3.1383709	0.4367214	2.2628205	5.2173913
pd1year	103	2.8751627	0.4875549	1.9642857	4.8913043
attachbase	130	2.1460753	0.7970523	0.8950617	5.0892857
attach1year	103	2.1013885	0.7718840	0.8653846	5.3043478

Form 3-2: The Frequency of the categorical variables

Table of trtgroup by smoker

trtgroup	smoker 0	smoker 1	Total
1	15 11.63 57.69 18.52	11 8.53 42.31 22.92	26 20.16
2	17 13.18 65.38 20.99	9 6.98 34.62 18.75	26 20.16
3	18 13.95 69.23 22.22	8 6.20 30.77 16.67	26 20.16
4	14 10.85 56.00 17.28	11 8.53 44.00 22.92	25 19.38
5	17 13.18 65.38 20.99	9 6.98 34.62 18.75	26 20.16
Total	81 62.79	48 37.21	129 100.00

Frequency Missing = 1

Table of trtgroup by race

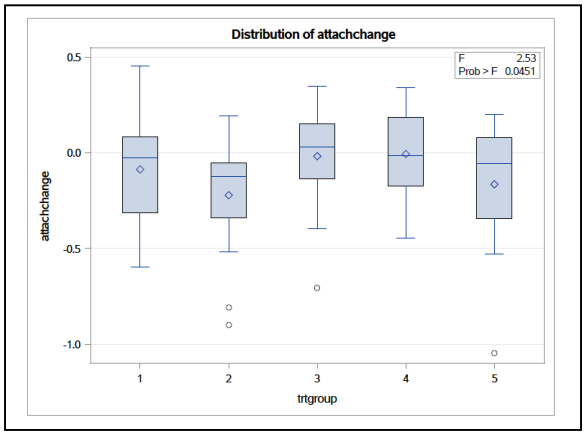
trtgroup	race 1	race 2	race 4	race 5	Total
1	0 0.00 0.00 0.00	2 1.54 7.69 22.22	1 0.77 3.85 33.33	23 17.69 88.46 20.18	26 20.00
2	1 0.77 3.85 25.00	1 0.77 3.85 11.11	1 0.77 3.85 33.33	23 17.69 88.46 20.18	26 20.00
3	1 0.77 3.85 25.00	5 3.85 19.23 55.56	0 0.00 0.00 0.00	20 15.38 76.92 17.54	26 20.00
4	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.77 3.85 33.33	25 19.23 96.15 21.93	26 20.00
5	2 1.54 7.69 50.00	1 0.77 3.85 11.11	0 0.00 0.00 0.00	23 17.69 88.46 20.18	26 20.00
Total	4 3.08	9 6.92	3 2.31	114 87.69	130 100.00

Table of trtgroup by sex

trtgroup	sex 1	sex 2	Total
1	11 8.46 42.31 20.37	15 11.54 57.69 19.74	26 20.00
2	10 7.69 38.46 18.52	16 12.31 61.54 21.05	26 20.00
3	11 8.46 42.31 20.37	15 11.54 57.69 19.74	26 20.00
4	11 8.46 42.31 20.37	15 11.54 57.69 19.74	26 20.00
5	11 8.46 42.31 20.37	15 11.54 57.69 19.74	26 20.00
Total	54 41.54	76 58.46	130 100.00

Generalized Linear Model (GLM) were applied to compare if the five different treatment groups would lead to significant changes in attach loss or pocket depth. Figure 1 showed there was significant difference for the five treatment groups for attach loss change (p=0.0451); but there was not significant difference for the five treatment groups for pocket depth change (p=0.0899 or p=0.1208).

Figure 1



Form4: t-test of treatment for attachchange

treatment	P-value	treatment	P-value
Trt1&2	0.0883	Trt2&3	0.0174
Trt1&3	0.3706	Trt2&4	0.0096
Trt1&4	0.2733	Trt2&5	0.5735
Trt1&5	0.3981		

Form5: GLM for attachchange and pdchange

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trtgroup	4	0.47831707	0.11957927	1.97	0.1063
attachbase	1	0.84801112	0.84801112	13.95	0.0003
race	1	0.10457025	0.10457025	1.72	0.1930
sex	1	0.02876959	0.02876959	0.47	0.4933
age	1	0.08468370	0.08468370	1.39	0.2410
smoker	1	0.14215054	0.14215054	2.34	0.1297
sites	1	0.01509983	0.01509983	0.25	0.6194

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trtgroup	4	0.52588496	0.13147124	1.88	0.1208
attachbase	1	0.00059475	0.00059475	0.01	0.9267
sex	1	0.27805833	0.27805833	3.98	0.0492
age	1	0.07995641	0.07995641	1.14	0.2878
race	1	0.06127884	0.06127884	0.88	0.3517
smoker	1	0.00395873	0.00395873	0.06	0.8125
sites	1	0.03928912	0.03928912	0.56	0.4555

T-test was further to investigate which treatment groups were significant different for the outcome, attachment loss change. Without adjustment, the average change of attach

loss was statistical significance (P=0.0451) among the five treatment groups, specifically between control group and low concentrations (Trt2&3: P=0.0174), or medium concentrations group (Trt2&4: P=0.0096). However, the Placebo group had not significant difference from all other groups. Adjusted by other factors, baseline, race, smoker, sex and age, by GLM shown in Form 5, the treatments were not significant anymore (p=0.1063). GLMselect also excluded the treatment groups in the model.

Conclusions:

The result showed that brushing the study gel on the gum with active ingredient, had no significant difference from the gel without any active ingredient. The new gel might not have clinical significance. After adjusted by baseline, race, smoker, sex and age, the treatments were not significant (p=0.1063) and lost statistical significance. The new gel did not lower the pocket depth since there was no significant difference for the five treatment groups with/without adjusted by other variables (p=0.1208 or p=0.0899).

Hence, adjusted by demographic factors, the new gel had no significant effect on the treatment population, and was not promising to test for clinical trial. But baseline may affect attach loss (p<.0001), and sex may affect pocket depth (p=0.033).

Limitation: The number of the precipitants was small, 26 samples in each treatment group. If the population increased, the model might be more accurate and the clinical significance might be much better addressed. Since the small group in the data, and it as hard to track the reason and conditions for the persons who failed to show up at one year, it was difficult to impute the accurate values for the missing data.

Reproducible research information: See below (**Reproducible Research**).