Final Project

Minerva University

CS130: Statistical Modeling: Prediction and Causal Inference

Prof. A. Diamond

April 19, 2023

Chirag Jeswani, Shivam Rawat, Tetiana Bas

Final Project

Executive Summary:

Due to the limitations of propensity score matching (PSM), genetic matching was proposed for improving covariate balance for a 2021 study by Hussein Aljawad et al. The propensity score matching results of the original paper were replicated and then the paper was extended by using genetic matching. Based on the results of the paper and our extension, genetic matching is suggested as the better method for balancing covariates used in the study.

Introduction:

The study titled “Three-dimensional evaluation of upper airway changes following rapid maxillary expansion: A retrospective comparison with propensity score matched controls” by Hussein Aljawad et al. (Public Library of Science, 2021) used propensity score matching (PSM) with the aim “to evaluate upper airway changes three-dimensionally following rapid maxillary expansion (RME) and compare the changes with matched controls” (Aljawad et al., 2021) because there is conflicting evidence related to the effect of RME treatment on upper airway dimensions (Zhao et al., 2010) (Iwasaki et al., 2013).

The main causal inference question of this study is to answer - ***what is the effect of RME on the upper airway across different dimensions*?** specifically airway volumes of nasopharynx and oropharynx and minimum cross-sectional areas (MCA) of oropharynx.

This is a causal inference question because the study evaluates if RME treatment “causes” changes across the aforementioned upper airway dimensions or not in patients with maxillary transverse deficiency.

To answer this question the authors of the study used 17 patients with MSE that were treated with RME as the case and 17 patients as the control (they were selected from a non-RME group of 33). PME was used to balance 5 covariates (age, gender, CBCT scan interval, sagittal skeletal pattern, and tongue posture) in the case and control group because this was an observational study (the assignment of patients to treatment was not random). Only MCA of the retropalatal segment showed statistically significant differences in upper airway dimensions after RME.

The dataset for replication was accessed from the Harvard Dataverse website. The website had two excel files, one with all the 17 cases and 33 controls with all the 5 covariate measurements for all the units and the second one with all the 17 cases and the 17 matched control with all the upper airway dimensions for all the units. Both the case and control data was collected from a hospital database with a comprehensive inclusion and exclusion criteria. Only 7 units out of 34 used for estimating the treatment effect were male, which limits the generalizability of this study to men because we can only make inferences about the matched data. There was no code for replication in the database.

Replication:

In the original study, the researchers used propensity scores matching to create a control group for their analysis of the effects of rapid maxillary expansion on the upper airway. Propensity score matching is a statistical method that aims to balance the distribution of covariates (i.e., potential confounding variables) between the treatment group and the control group. The goal is to create a control group that is as similar as possible to the treatment group, except for the treatment itself, so that any differences observed between the groups can be attributed to the treatment.

To create the control group, the researchers first identified a group of patients who had undergone rapid maxillary expansion. They then identified a group of patients who had not undergone this treatment but were similar to the treatment group in terms of age, sex, and other relevant factors that might influence the outcome. The propensity score was calculated for each patient in both groups based on their baseline characteristics. This score reflects the probability of each unit being assigned to the treatment group, given a set of baseline characteristics.

The researchers then matched each patient in the treatment group with one or more patients in the control group who had a similar propensity score. This matching was done using a nearest-neighbor algorithm, which selects the closest control group member(s) to each patient in the treatment group based on their propensity score.

After matching, the distribution of baseline characteristics was compared between the treatment and control groups to ensure that they were similar. This was done using standardized mean differences, which measure the difference in means between the treatment and control groups relative to the pooled standard deviation.

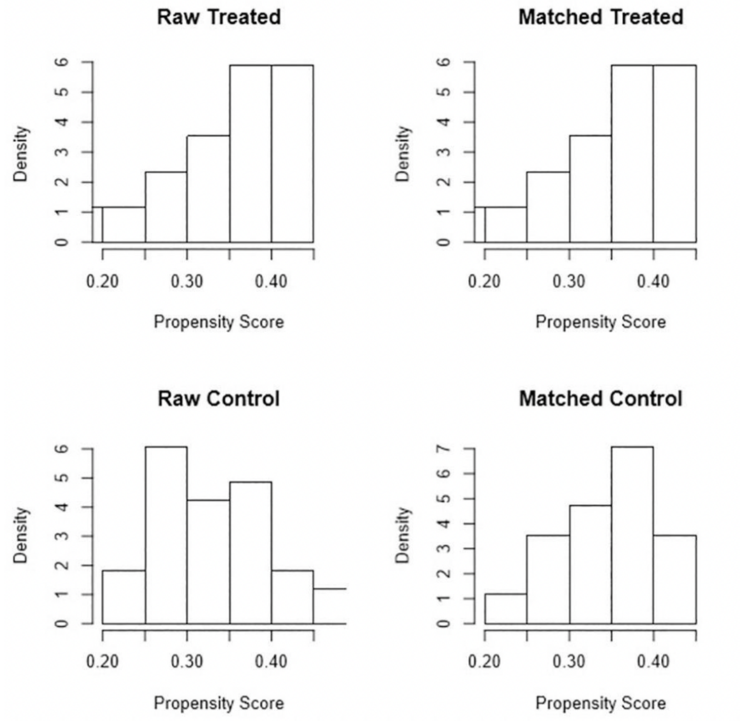


Figure 1: Histograms showing the density of propensity score distribution in the treated and control groups before and after matching. This is the ***original figure*** *from the study (Aljawad et al., 2021).*

As we can see from figure 1, the distributions of the matched control and matched treated variables are more similar than before matching, indicating that the matching process was successful in reducing selection bias. Since the PSM was performed on the treated units, the distribution of the treatment group remained the same pre and post treatment.

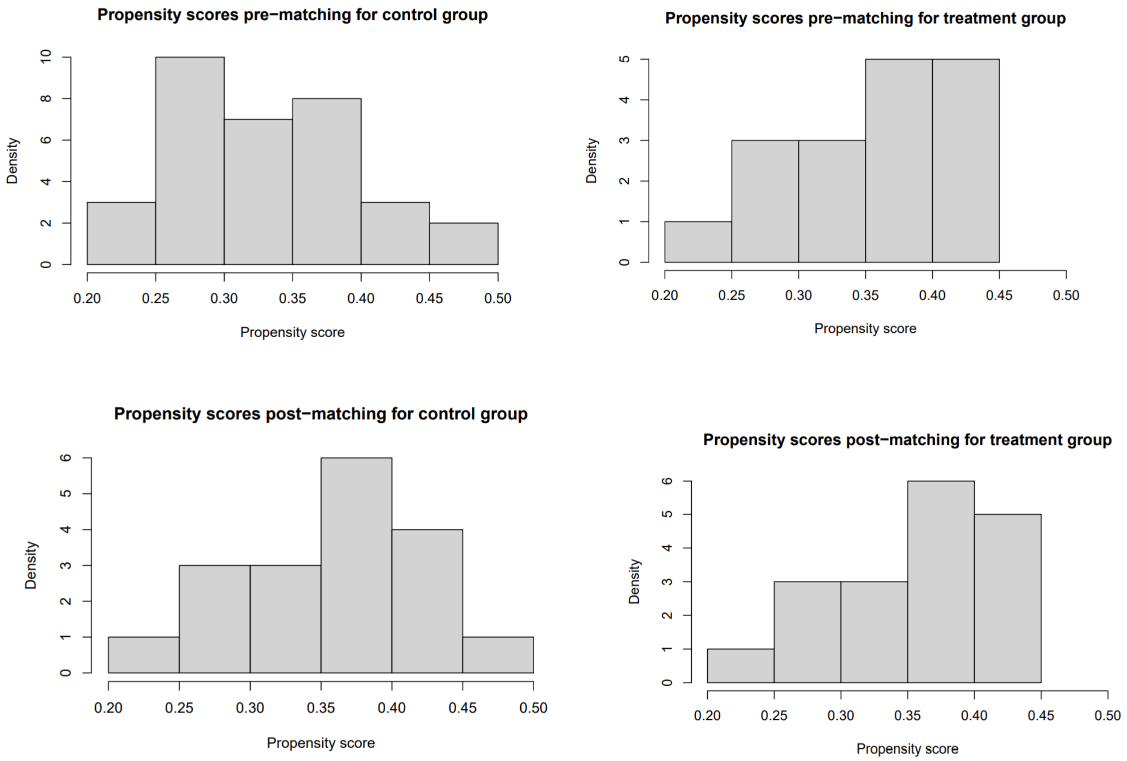


Figure 2: Histograms showing the density of propensity score distribution in the treated and control groups before and after matching. This is the replicated figure of the original figure 1

In order to replicate the plots, we first calculated the propensity scores by performing logistic regression on covariates (in this case: age, gender, tongue posture etc.), then, we saved all the scores separately after that match balance was run, to find the optimal control-treatment pairs. The propensity scores from this represent the post-matching histograms.

We know that the matching process worked well because the post-matching control and treatment graphs look more similar than pre-matching ones. This means that the distribution of covariates is balanced and they would ideally not influence our analysis.

There are several potential disadvantages of using propensity score matching in this study:

1. Limited sample size: The study had a small sample size of only 17 patients, which may limit the generalizability of the findings.
2. Overfitting: The paper does not mention overfitting as a concern with PSM analysis. However, it is important to note that overfitting can be a potential issue with PSM, if too many covariates are used for matching. This can lead to a reduction in the accuracy of the estimates.
3. Unmeasured confounding: PSM can only control for measured confounding variables, and there may be unmeasured confounding variables that could affect the results.

The paper does not explicitly address unmeasured confounding variables. However, the authors attempted to control for potential confounding variables that might affect nasopharyngeal and oropharyngeal airway dimensions by including several covariates in the propensity score matching (PSM) analysis. The covariates used for matching included age, gender, CBCT scan interval, sagittal skeletal pattern, and tongue posture. By including these covariates in the matching process, the authors attempted to reduce the impact of potential confounding variables on the results. However, it is important to note that PSM can only control for measured confounding variables, and there may be unmeasured confounding variables that could affect the results. Therefore, while the authors attempted to control for potential confounding variables using PSM, it is possible that there are unmeasured confounding variables that could affect the results.

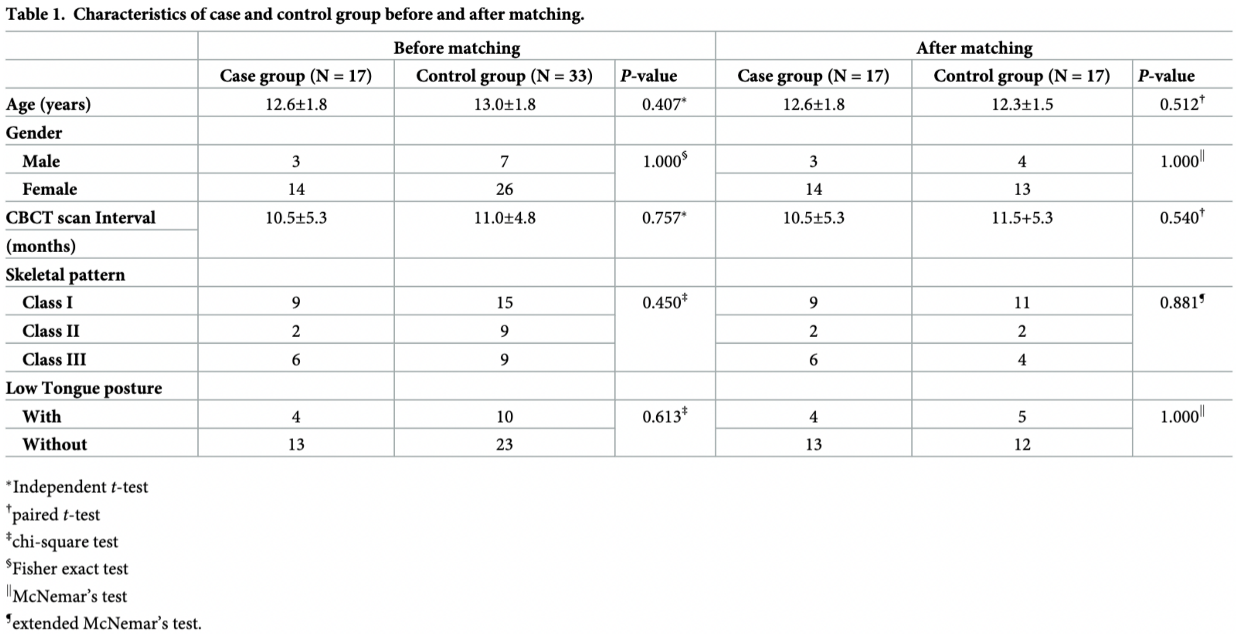


Table 1. Characteristics of case and control group before and after matching.

	Before matching			After matching		
	Case group (N = 17)	**Control group (N = 33)**	***P*-value**	**Case group (N = 17)**	**Control group (N = 17)**	***P*-value**
Age (years)	12.6±1.8	13.0±1.8	0.407[*]	12.6±1.8	12.3±1.5	0.512[†]
Gender						
Male	3	7	1.000[§]	3	4	1.000[‖]
Female	14	26		14	13	
CBCT scan Interval (months)	10.5±5.3	11.0±4.8	0.757[*]	10.5±5.3	11.5+5.3	0.540[†]
Skeletal pattern						
Class I	9	15	0.450[‡]	9	11	0.881[¶]
Class II	2	9		2	2	
Class III	6	9		6	4	
Low Tongue posture						
With	4	10	0.613[‡]	4	5	1.000[‖]
Without	13	23		13	12	

[*] Independent *t*-test
[†] paired *t*-test
[‡] chi-square test
[§] Fisher exact test
[‖] McNemar's test
[¶] extended McNemar's test.

Table 1: The results of the evaluation of the balance of covariates between the case and control groups before and after propensity score matching (PSM) This table is taken from the original study (Aljawad et al., 2021)

Table 1 shows the means and standard deviations of the covariates for the case and control groups before and after matching, as well as the p-values for the differences between the groups. The table indicates that before matching, there were significant differences between the case and control groups in age, gender, CBCT scan interval, sagittal skeletal pattern, and tongue posture. However, after matching, there were no significant differences between the groups in any of these covariates, indicating that the PSM was successful in creating a more comparable case and control group. The evaluation of the balance of covariates before and after matching is an important step in assessing the quality of PSM. The fact that the PSM was able to balance the covariates between the case and control groups suggests that the matching process was successful in reducing selection bias and creating a more comparable case and control group.

Extension:

Propensity score matching may not always achieve perfect balance between treated and control groups partially due to the reasons explained before. One way to improve propensity score matching is to perform genetic matching, which is a powerful technique that uses a genetic algorithm to find optimal weights for each covariate variable. Genetic matching can improve balance by taking into account the interdependence between covariate variables, which may not be fully captured by the propensity score model. In genetic matching, the algorithm searches for weights that minimize the distance between treated and control groups, based on a set of genetic markers or other relevant variables. This can result in better balance and a more accurate estimate of the treatment effect.

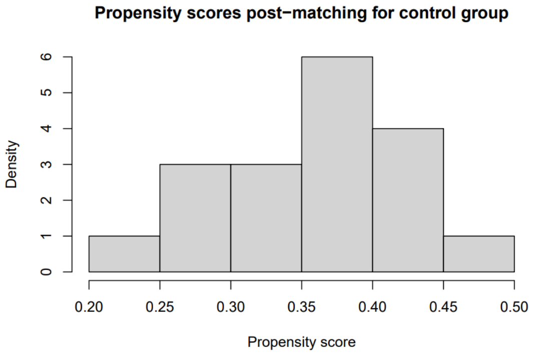
Metric	**Propensity score**		**Genetic matching**		**Without matching**	
Group	Treated	Control	Treated	Control	Treated	Control
Mean Age	151.3889	150.9444	152.2222	152.5556	151.2353	156.4848
Mean CBCT	10.38889	9.944444	10.16667	10.27778	10.52941	11
Gender Female	15	15	15	15	14	26

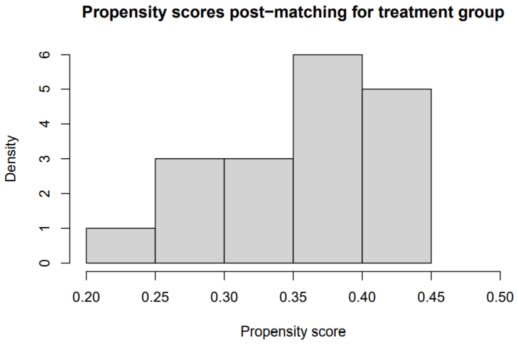
Male	3	3	3	3	3	7
Sagittal Class1	10	10	9	10	9	15
Class 2	2	3	3	3	2	9
Class 3	6	5	6	5	6	9
Tongue without	14	14	14	14	13	23
Tongue with	4	4	4	4	4	10

Table 2: This is the replication of table 1 with extension data from genetic matching. The results of the evaluation of the balance of covariates between the case and control groups before any method was applied, after propensity score matching (PSM) and after genetic matching

Table 2 successfully replicated table 1 in terms of getting the same covariate values for treatment and control groups for the before matching and after propensity score matching. To improve this, we added an additional column representing the values for the covariates after genetic matching was performed.

Based on table 2 we can see the comparison in the balance of the covariates for treatment and control groups. The main covariate variables in this case include mean age, mean CBCT, gender, sagittal class, and tongue position. The table shows that both propensity scores matching and genetic matching were effective in balancing the covariate variables compared to the results obtained without matching. However, genetic matching was able to achieve even better balance, particularly for mean age and mean CBCT. This suggests that genetic matching reduced the imbalance in these covariates better than propensity score matching.Therefore, while propensity scores matching was able to improve the balance of the covariate variables, genetic matching was even more effective in achieving balance, as shown by the improved balance in almost all covariates and an especially big improvement for mean age and mean CBCT since we have the same means for both treatment and control groups after genetic matching was applied. Overall, this indicates that genetic matching proved to be a more powerful technique for reducing bias and achieving balance for this study.





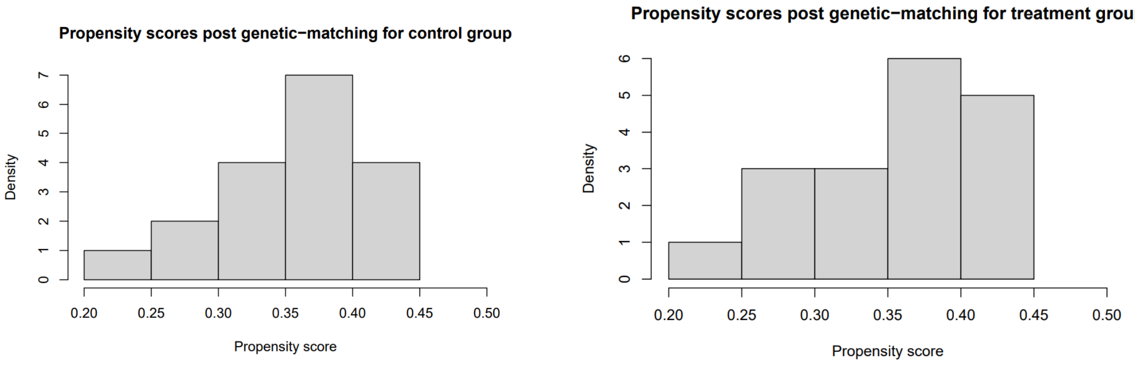


Figure 3: The figure represents the collection of histograms which compares how genetic and prop score matching performed. The first row has prop scores, the second has genetics.

	Propensity score matching		*Genetic matching*		*Difference in p-value (GM - PSM)*
	Before matching	*After matching*	*Before matching*	*After matching*	
Age	*0.76746*	*1*	*0.7674*	*1*	*-*
Gender	*0.40895*	*0.7805*	*0.40895*	*0.76687*	*-0.013663*
CBTB scan interval (months)	*0.76164*	*0.65209*	*0.76164*	*0.953*	*+0.30091*
Skeletal pattern	*0.98354*	*0.31798*	*0.98353*	*0.82118*	*+0.5032*
Low tongue posture	*0.61536*	*1*	*0.61536*	*1*	*-*
Minimum p-value	*0.40895*	***0.31798***	*0.40895*	***0.76687***	

Table3: p-values for every covariate for PSM and GM (before and after matching)

When comparing the replicated distributions for propensity score matching and genetic matching in Figures 2 and 3, it is evident that genetic matching produces more similar control and treatment group distributions than propensity score matching. This is a positive indication that genetic matching may yield better results in achieving balance between the groups.

However, visual comparison of distributions alone is not sufficient to draw conclusions about the efficiency of matching techniques. It is also important to consider statistical measures such as p-values to assess the performance of the methods. In this case, we can assess the efficiency of the matching techniques by looking at the p-values based on the t-test of the variables. Higher p-values indicate better balance between the treatment and control groups because it indicates no statistical difference in distribution of a covariate in the treatment and control group.

The minimum p-value after matching for genetic matching is significantly bigger than the minimum p-value after PSM.Together with comparison of p-values (See Table 3) confirms our conclusion from the histograms that genetic matching was better in terms of improving balance between covariates.

Difference in balance can change the results of ATE and hence our inferences about the causal inference question we are trying to answer.

Metric	**Propensity Scores Matching**		**Genetic Matching**	
	Treatment effect	**p-value**	**Treatment effect**	**p-value**
Nasal	0.6882353	0.1739	0.4647059	0.3483
RetroPlattal V	1.205882	0.08149	1.258824	0.08302

RetroPlattal MCA	54.94706	0.03504*	52.46471	0.06388
RetroGlossal V	1.7	0.1023	2.282353	0.01978*
RetroGlossal MCA	46.1588	0.06377	53.97059	0.03379*

Table 4: Table represents the comparison of the Average treatment effect for genetic matching and the propensity scores. (*) indicate statistically significant treatment effect.

From table 4 we can see that there are not any substantial differences in ATE after for PSM groups and GM groups. These results follow from the results in the original study and thus supports the findings of the original paper. Only the retroglossal segment of oropharynx (V) had considerable difference in ATE.

We can note the appearance of statistical significance in two outcome variables and disappearance of statistical significance in one outcome variable after genetic matching (See Table 4). Based on this genetic matching provides better evidence that RME has a positive treatment effect on the upper airway.

Conclusion:

Based on these results we can conclude that genetic matching was a better technique for matching in the context of this specific study because it leads to better balance in covariates and leads to more statistically significant results.

AI statement: We used the AI tool Humata to help us navigate the paper and Chat Gpt to help us make our writing more coherent.

References

Aljawad, H., Lee, K.-M., & Lim, H.-J. (2021). Three-dimensional evaluation of upper airway changes following rapid maxillary expansion: A retrospective comparison with propensity score matched controls. *PLoS ONE*, *16*(12), e0261579. https://doi.org/10.1371/journal.pone.0261579

Iwasaki, T., Saitoh, I., Takemoto, Y., Inada, E., Kakuno, E., Kanomi, R., Hayasaki, H., & Yamasaki, Y. (2013). Tongue posture improvement and pharyngeal airway enlargement as secondary effects of rapid maxillary expansion: A cone-beam computed tomography study. *American Journal of Orthodontics and Dentofacial Orthopedics: Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, *143*(2), 235–245. https://doi.org/10.1016/j.ajodo.2012.09.014

Zhao, Y., Nguyen, M., Gohl, E., Mah, J. K., Sameshima, G., & Enciso, R. (2010). Oropharyngeal airway changes after rapid palatal expansion evaluated with cone-beam computed tomography. *American Journal of Orthodontics and Dentofacial Orthopedics: Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, *137*(4 Suppl), S71-78. https://doi.org/10.1016/j.ajodo.2008.08.026

Appendix A: Contribution Statement

Chirag: Mainly responsible for coding and programming, as well as contributing to writing the explanations of the process. Also edited and revised the paper.

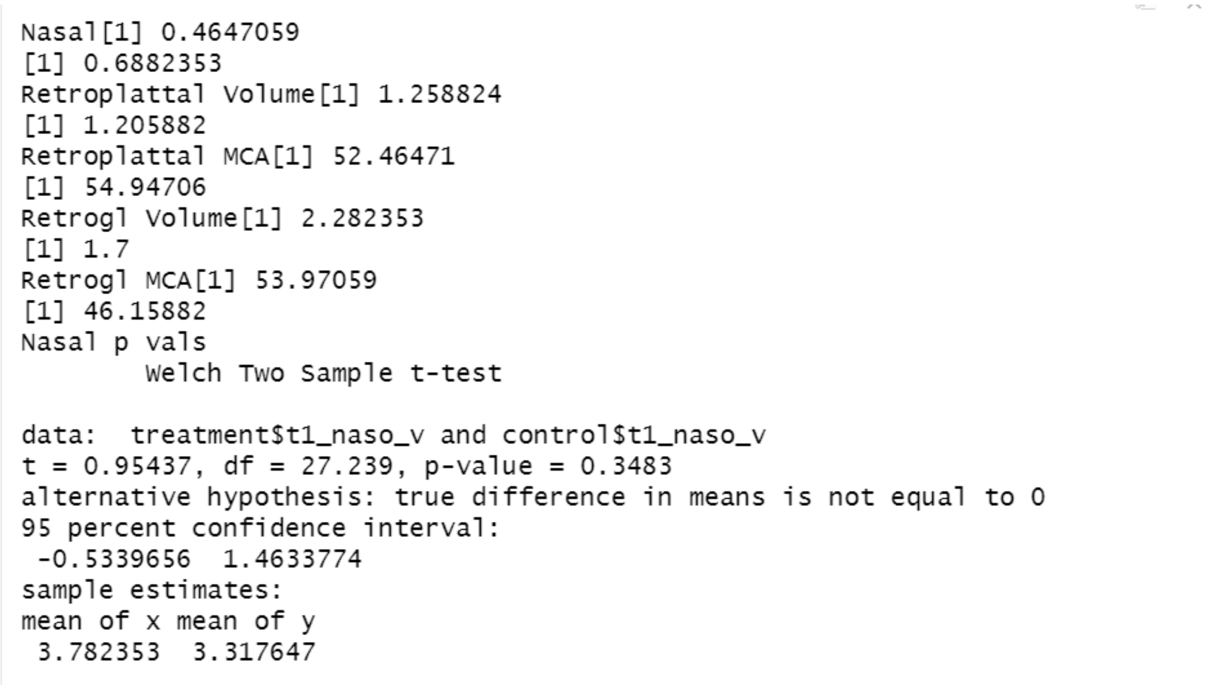
Tanya: Produced all the tables, analyzed them. Focused on writing the section for replication and extension, and also helped with editing and revising the paper.

Shivam: Found the research paper to replicate, wrote the executive summary, introduction, and improved conclusion and extension. Additionally, he contributed to editing and revising the paper and communication between members.

Link to the code:

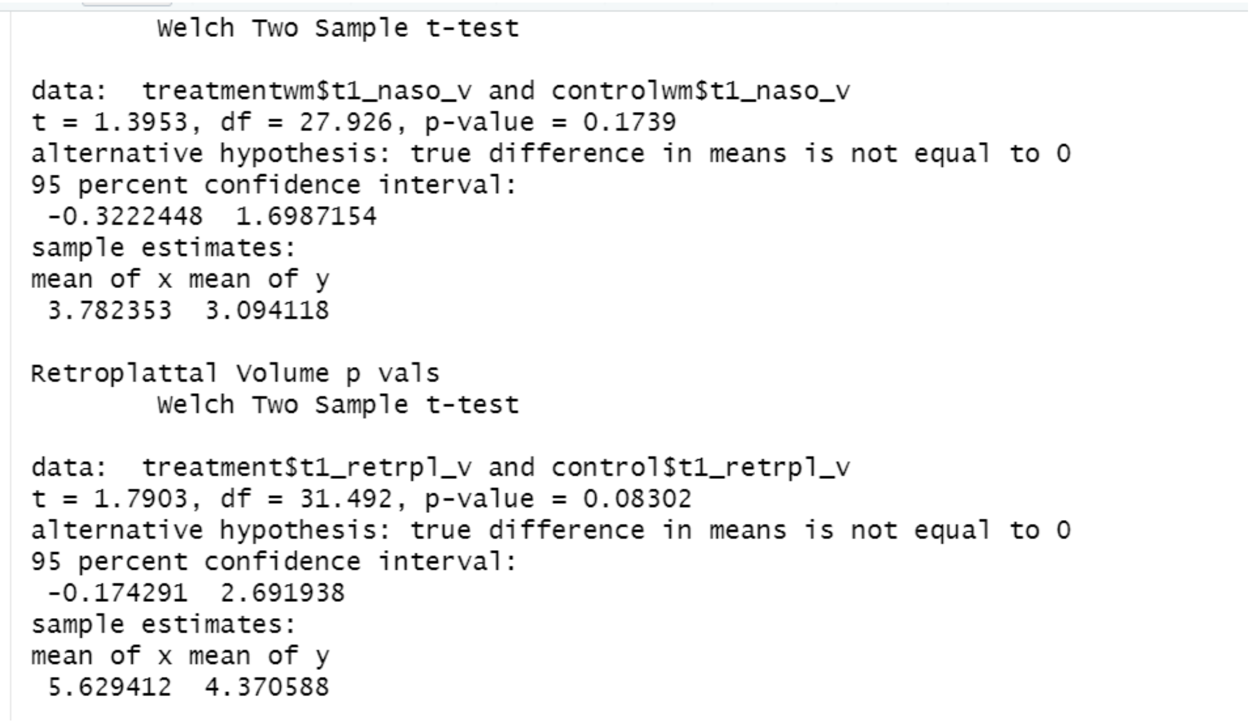
https://drive.google.com/drive/folders/1X9x38eiBUMT197AFelD-nQOXPlqEvK8r?usp=sharing

Appendix B



```
Nasal[1] 0.4647059
[1] 0.6882353
Retroplattal Volume[1] 1.258824
[1] 1.205882
Retroplattal MCA[1] 52.46471
[1] 54.94706
Retrogl Volume[1] 2.282353
[1] 1.7
Retrogl MCA[1] 53.97059
[1] 46.15882
Nasal p vals
        Welch Two Sample t-test

data:  treatment$t1_naso_v and control$t1_naso_v
t = 0.95437, df = 27.239, p-value = 0.3483
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5339656  1.4633774
sample estimates:
mean of x mean of y
 3.782353  3.317647
```

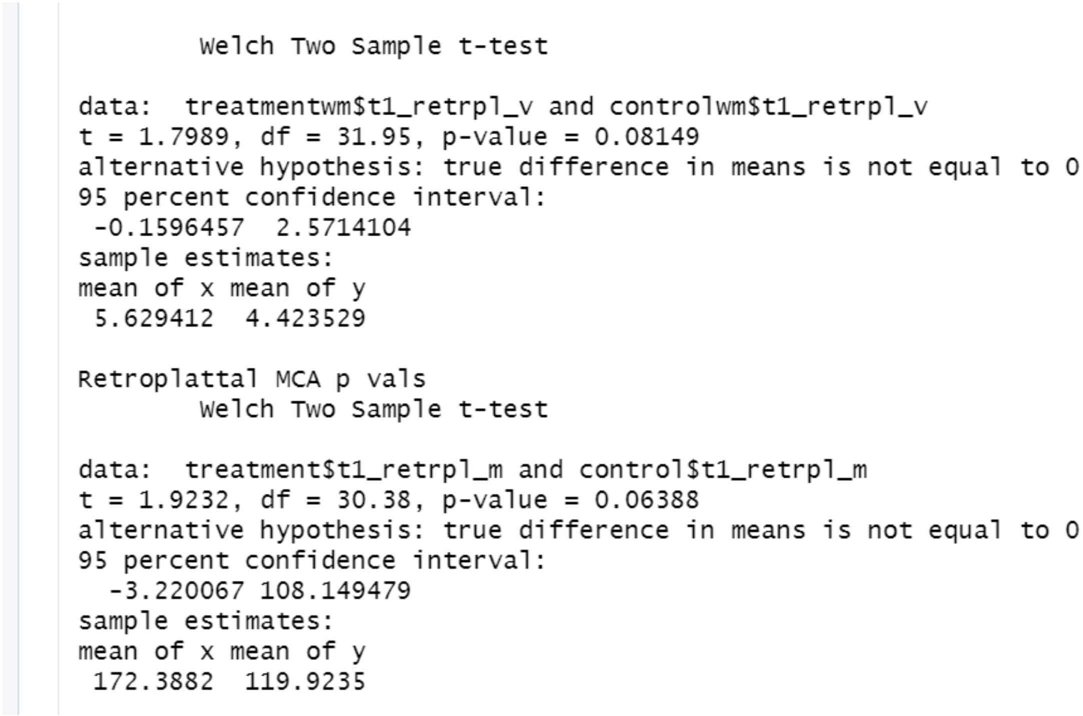


```
        Welch Two Sample t-test

data:  treatmentwm$t1_naso_v and controlwm$t1_naso_v
t = 1.3953, df = 27.926, p-value = 0.1739
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3222448  1.6987154
sample estimates:
mean of x mean of y
 3.782353  3.094118

Retroplattal Volume p vals
        Welch Two Sample t-test

data:  treatment$t1_retrpl_v and control$t1_retrpl_v
t = 1.7903, df = 31.492, p-value = 0.08302
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.174291  2.691938
sample estimates:
mean of x mean of y
 5.629412  4.370588
```

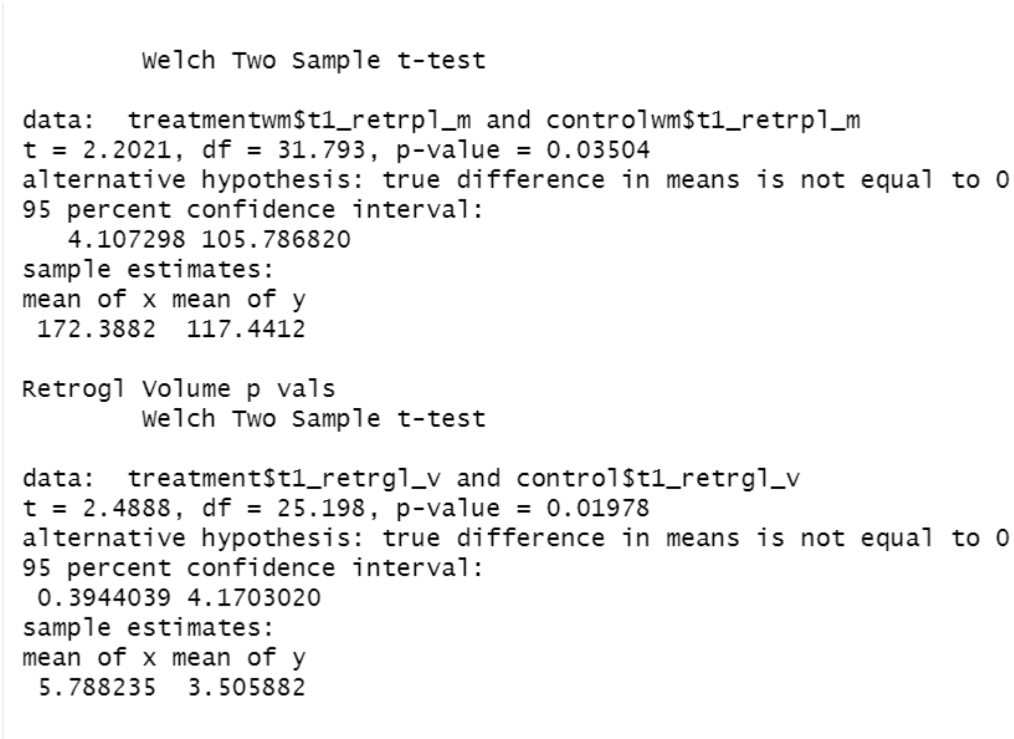


```
        Welch Two Sample t-test

data:  treatmentwm$t1_retrpl_v and controlwm$t1_retrpl_v
t = 1.7989, df = 31.95, p-value = 0.08149
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1596457  2.5714104
sample estimates:
mean of x mean of y 
 5.629412  4.423529 

Retroplattal MCA p vals
        Welch Two Sample t-test

data:  treatment$t1_retrpl_m and control$t1_retrpl_m
t = 1.9232, df = 30.38, p-value = 0.06388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.220067 108.149479
sample estimates:
mean of x mean of y 
 172.3882  119.9235 
```

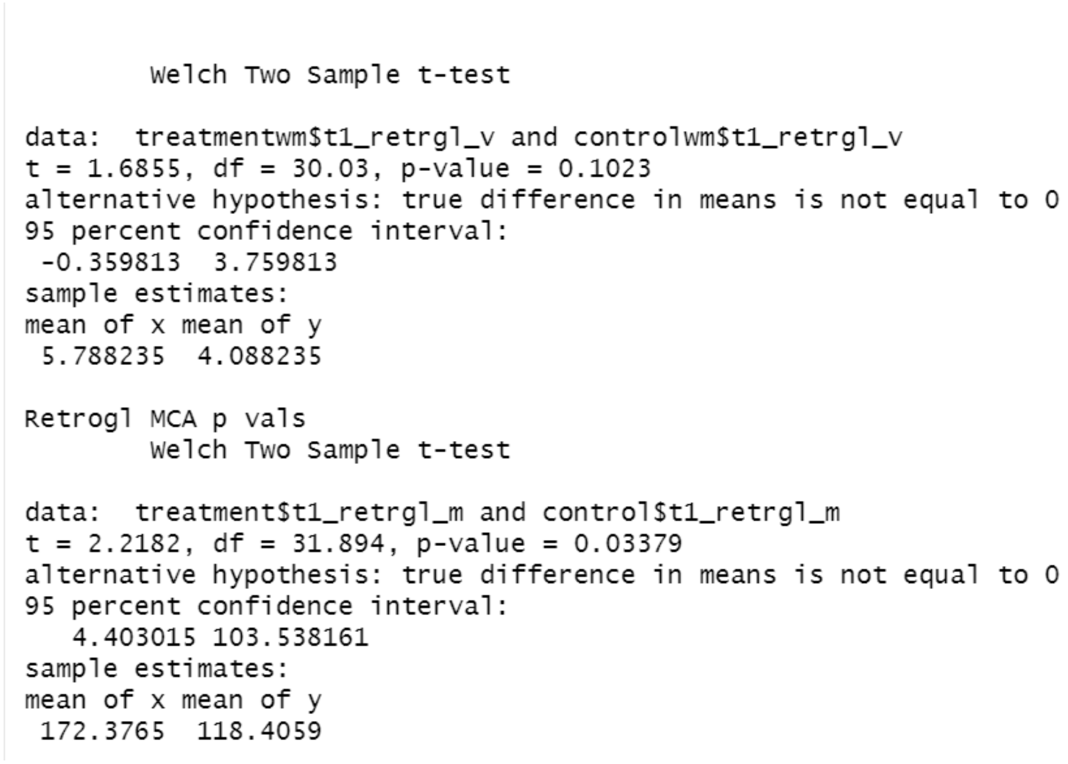


```
        Welch Two Sample t-test

data:  treatmentwm$t1_retrpl_m and controlwm$t1_retrpl_m
t = 2.2021, df = 31.793, p-value = 0.03504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   4.107298 105.786820
sample estimates:
mean of x mean of y 
 172.3882  117.4412 

Retrogl Volume p vals
        Welch Two Sample t-test

data:  treatment$t1_retrgl_v and control$t1_retrgl_v
t = 2.4888, df = 25.198, p-value = 0.01978
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3944039 4.1703020
sample estimates:
mean of x mean of y 
 5.788235  3.505882 
```

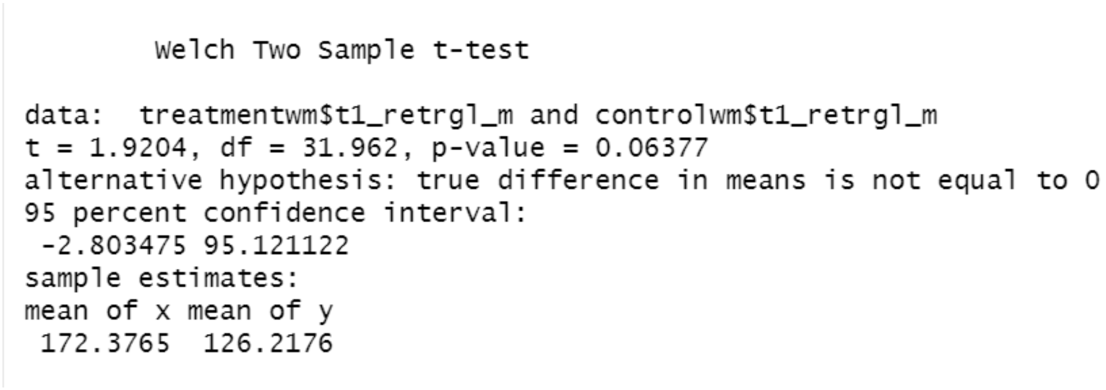


```
        Welch Two Sample t-test

data:  treatmentwm$t1_retrgl_v and controlwm$t1_retrgl_v
t = 1.6855, df = 30.03, p-value = 0.1023
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.359813  3.759813
sample estimates:
mean of x mean of y 
 5.788235  4.088235 

Retrogl MCA p vals
        Welch Two Sample t-test

data:  treatment$t1_retrgl_m and control$t1_retrgl_m
t = 2.2182, df = 31.894, p-value = 0.03379
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   4.403015 103.538161
sample estimates:
mean of x mean of y 
 172.3765  118.4059 
```



```
        Welch Two Sample t-test

data:  treatmentwm$t1_retrgl_m and controlwm$t1_retrgl_m
t = 1.9204, df = 31.962, p-value = 0.06377
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.803475 95.121122
sample estimates:
mean of x mean of y 
 172.3765  126.2176 
```