Bayer Take Home Project Presentation

Siyi Wei

05/07/2021

Data Description

In this dataset, we have 444 observations with totally 9 variables. Where 4 of them are categorical and 5 of them are numerical. Our controlled variable is the assigned treatment (arm), which is a non balanced binary variable with 221 PLACEBOs and 223 ACTIVEs. More details could be viewed below:

	subject	nosebleeds	duration	arm
Categorical	-	-	-	Binary
Numerical	Discrete	Discrete	Discrete	-

	country	eye.colour	tissue.use	previous.year	mucus.viscosity
Categorical	Multivariate	Multivariate	Binary	-	-
Numerical	-	-	-	Discrete	Continuous

Feature Engineering

Data Aggregation

There are three datasets provided for this analysis. Efficacy, Randomization and Subject.

	Efficacy	Randomization	Subject
Observations	444	444	444
Features	3	2	6
UID	subject	subject	subject
Feature Name	nosebleeds, durantion	arm	country, eye.colour, tissue.use, previous.year, mucus.viscosity

All of them have a unique identifier "subject". Which does not contain any duplicates. Our first step is to aggregate three tables based on this unique identifier "subject".

Impute the NULL values

We found there are multiple NA entries in eye.colour. Since there is no information on why those entries are missing. Based on simple inference, geographical position and race are the key factors that relate to eye colours. Therefore we decide to impute the NA values in eye.colour based on the most common eye.colour in the same country. However, country "G" only has one eye colour and the entry is missing. So we set it to "BLACK" randomly.

There is also one patient that has a NA value for mucus.viscosity. We will use the mean of mucus.viscosity based on other patients in the same tissue.use level.

Change the unit of nosebleeds to times/year

We could see for different patients, the time that the subject was on the study (duration) are different. This also affects the number of nosebleeds observed in study (nosebleeds). To construct a better metric which could measure the treatment effect. We decide to transform nosebleeds unit to times/year by dividing the current duration and times 365 days. Then round the new nosebleeds value to its nearest integers.

Construct measure of treatment effect

This new treatment (superdupripine) was produced for the unmet clinical need of recurrent serious nosebleeds. Therefore, the best measure for the treatment effect would be the times of nosebleeds reduced from last year.

Exploratory Data Analysis

Data Statistics

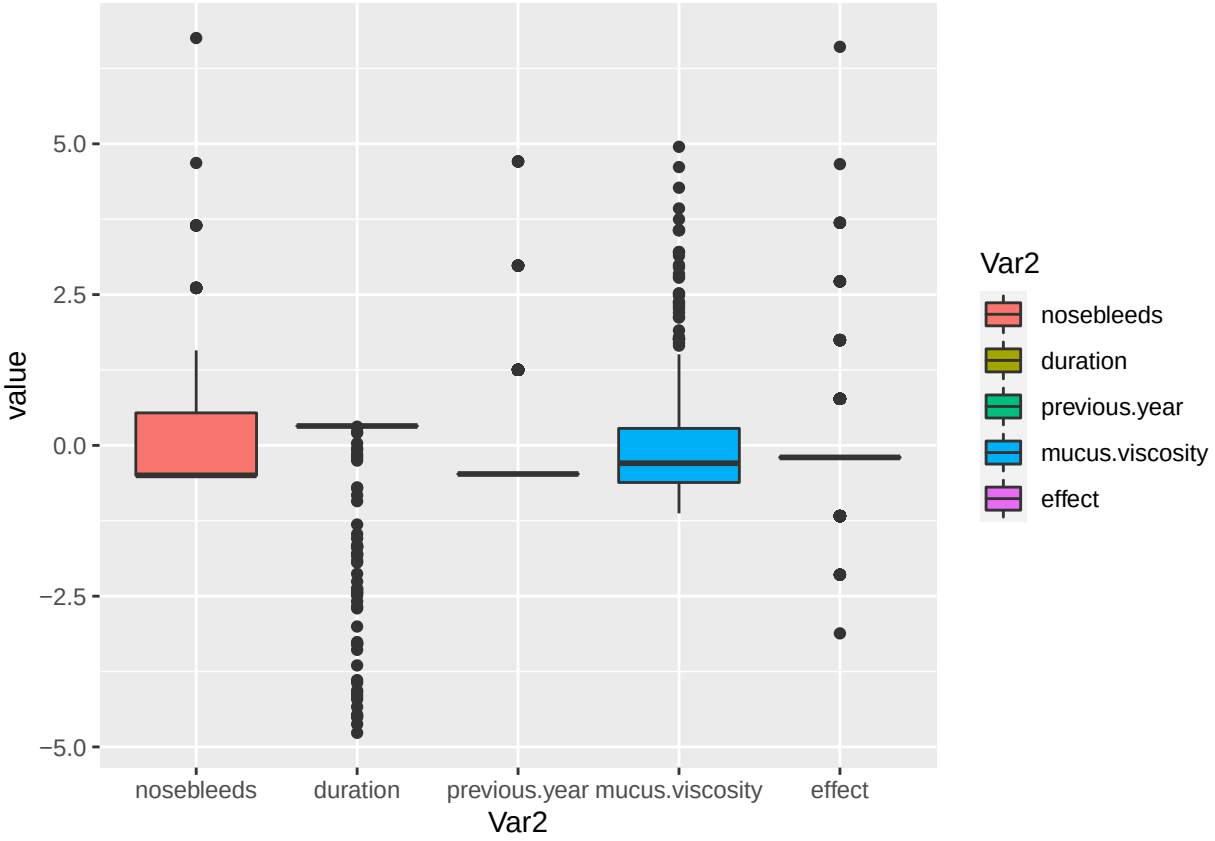
The Summary of Statistics could be viewed below:

```
##    nosebleeds        duration          arm         country   eye.colour
##  Min.   :0.0000   Min.   : 10.0   ACTIVE :223   H      :85   BLACK: 54
##  1st Qu.:0.0000   1st Qu.:365.0   PLACEBO:221   B      :72   BLUE :335
##  Median :0.0000   Median :365.0                 A      :50   BROWN: 55
##  Mean   :0.4797   Mean   :342.6                 D      :48
##  3rd Qu.:1.0000   3rd Qu.:365.0                 E      :47
##  Max.   :7.0000   Max.   :365.0                 F      :44
##                                                 (Other):98
##   tissue.use  previous.year   mucus.viscosity      effect
##  HIGH  :198   Min.   :2.000   Min.   :0.0000   Min.   :-5.000
##  MEDIUM:246   1st Qu.:2.000   1st Qu.:0.6152   1st Qu.:-2.000
##               Median :2.000   Median :1.0000   Median :-2.000
##               Mean   :2.275   Mean   :1.3566   Mean   :-1.795
##               3rd Qu.:2.000   3rd Qu.:1.6957   3rd Qu.:-2.000
##               Max.   :5.000   Max.   :7.3174   Max.   : 5.000
##
```

Scaled Numerical Data Distribution

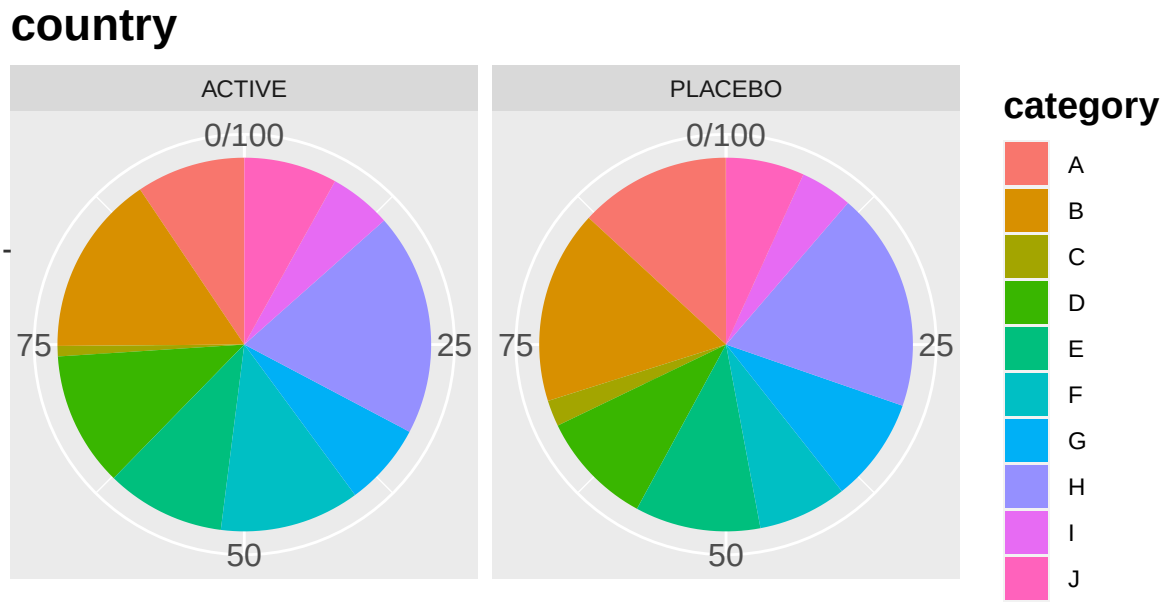
From the numerical distribution plot. We could see there are huge amounts of outliers inside of our dataset. We will address their details below

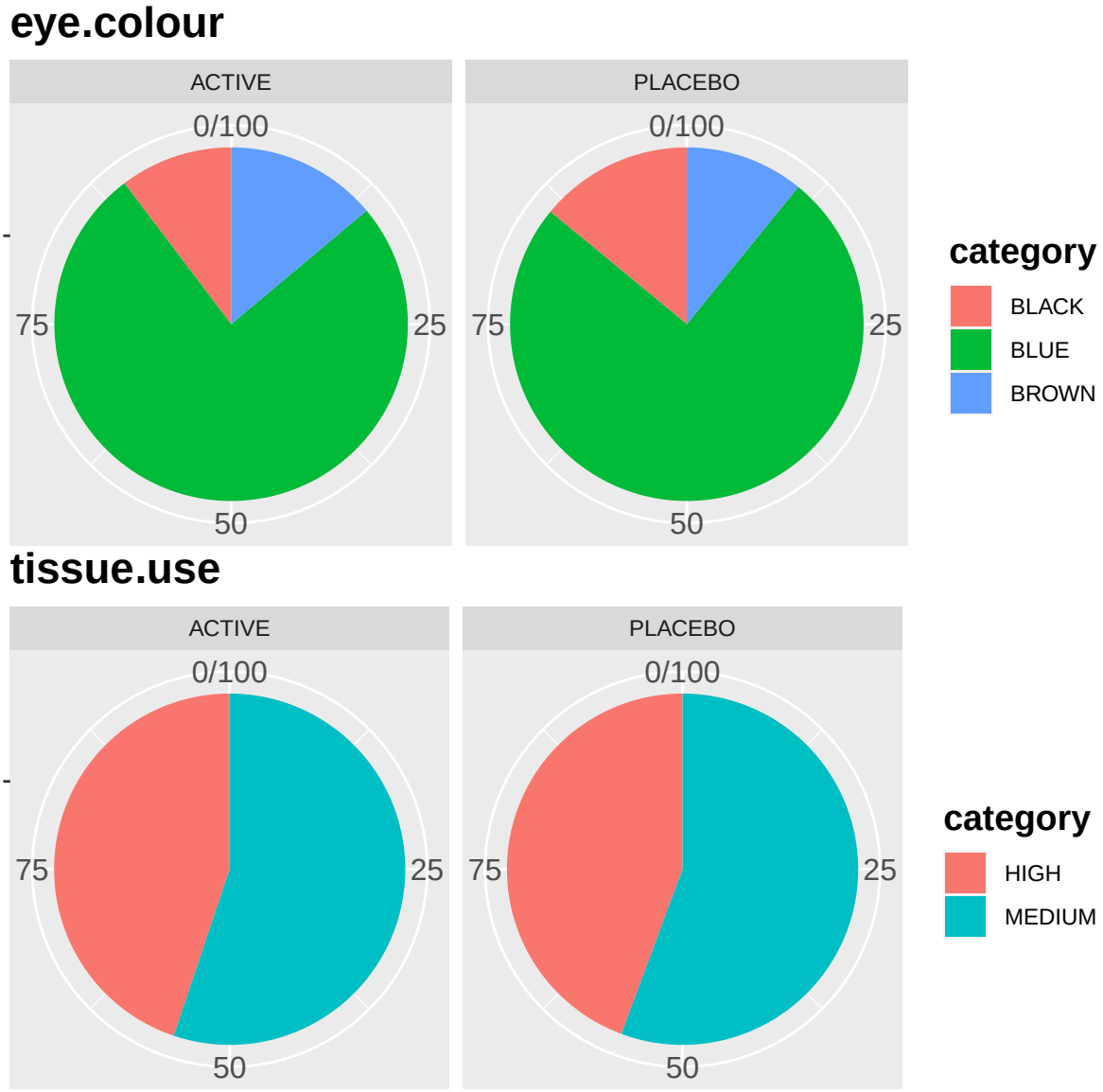
- nosebleeds & previous year & mucus.viscosity: Since there are some specific patients who have more serious nosebleeds than the others. We need to investigate more on those valuable outliers.
- duration: As we have explained, the duration of the patients in study varies a lot. But most of the durations are close to one year. So we have corrected all the patients' duration and its related features (nosebleeds) to one year.
- effect: The effect of the treatments heavily depends on the patient's nosebleeds level in the past year. From the plot we could see that patients' conditions are worse (positive outliers) or better (negative outliers). We also need to investigate those outliers.



Categorical Data Visualization

From the categorical data visulization, we could confirm the balance between placebo group and active group on country, eye.colour and tissue.use. There should not be any sigficant fairness problems caused by groupwise differences.



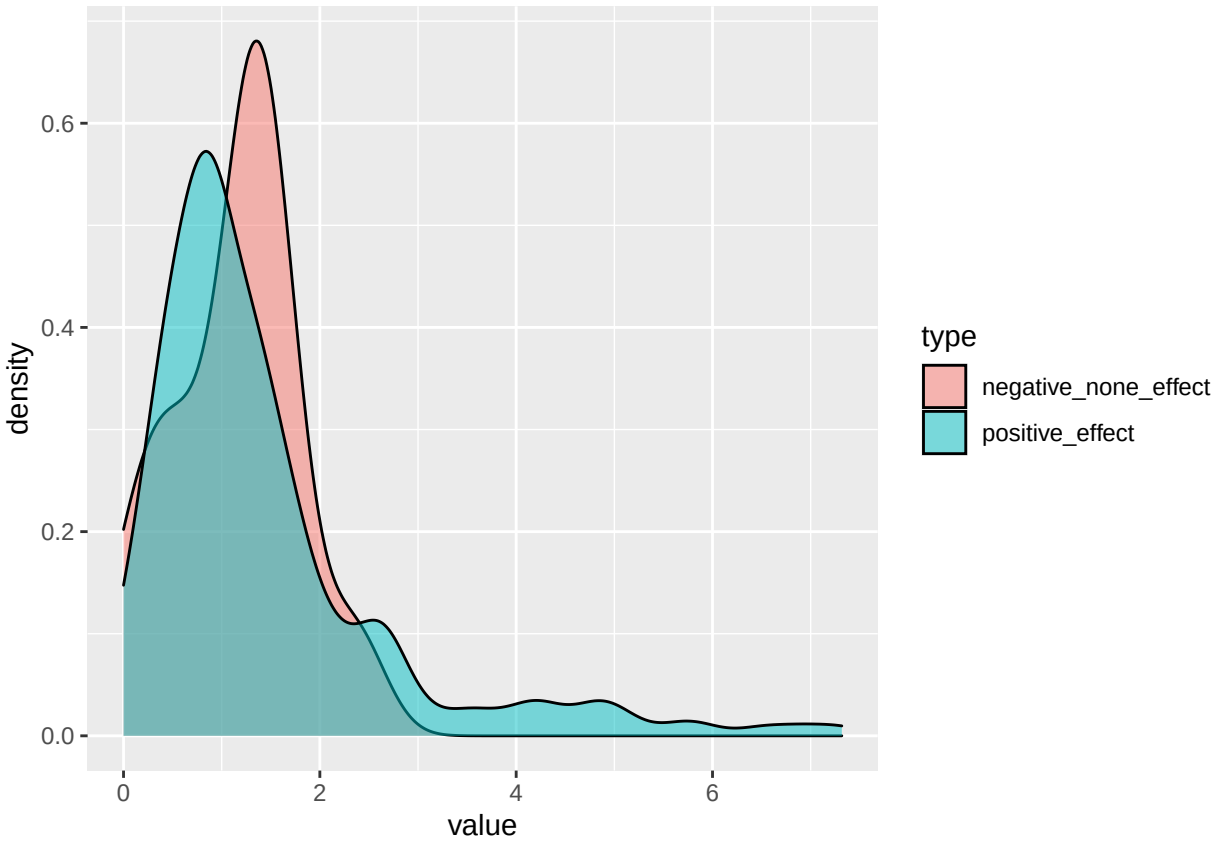


Data Values

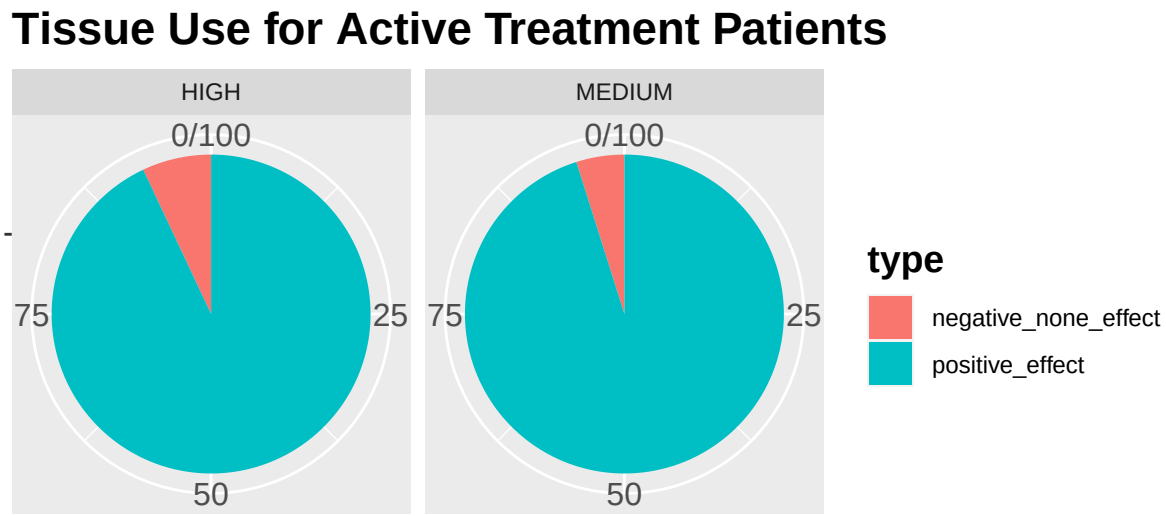
How might you show how the treatment effect depends on nasal mucus viscosity? What about the effect of paper tissues?

To demonstrate the treatment effect indeed depends on nasal mucus viscosity. We need to refer to the continuous distribution plot we had in EDA. Since we already know there are outliers for both positive and negative sides. **We want to know for the positive side outliers with active treatment effects, where patients have more nosebleeds than last year. What are those patients' nasal mucus viscosity distributions? Then we could compare them with the negative outliers with the same condition**

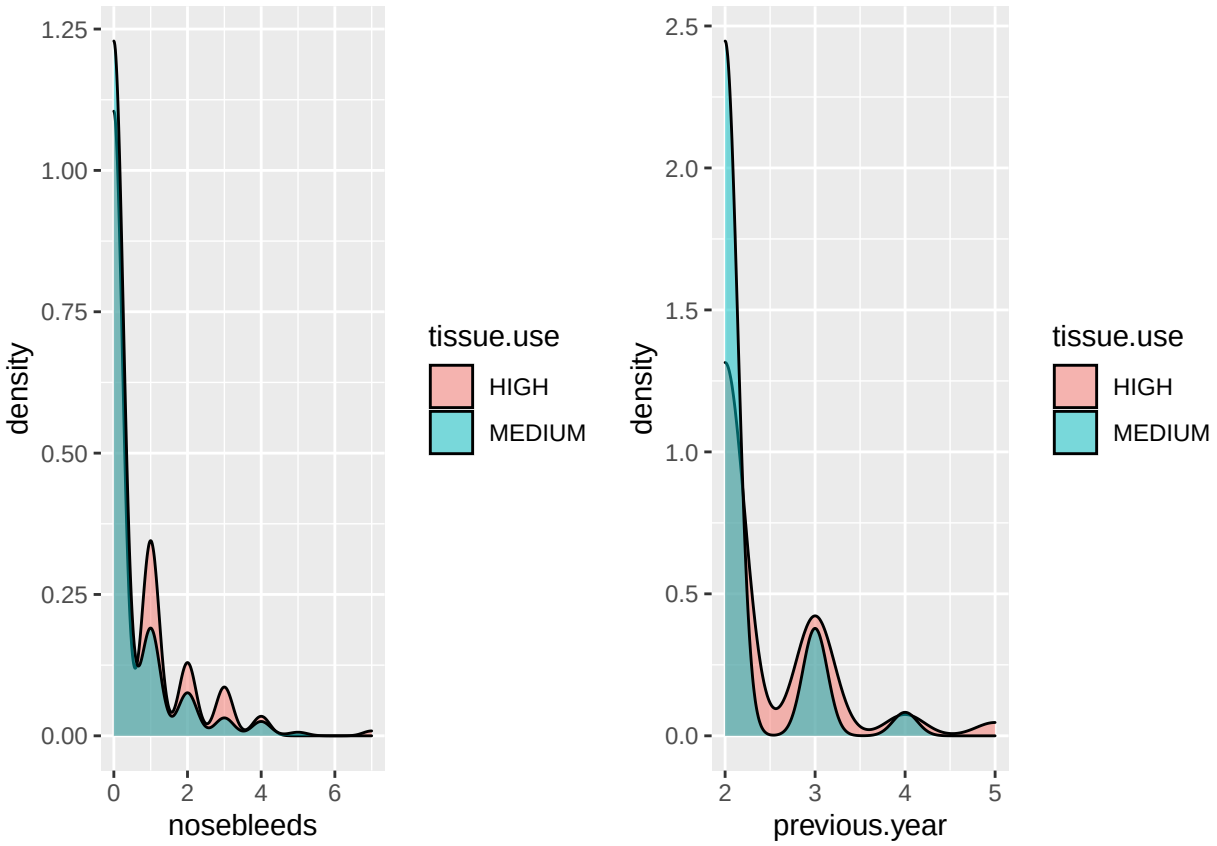
From the graph below. We could see the active treatment group. Even though the peak of mucus viscosity for the negative/none effect group and the positive effect group are close. Positive effect group has a heavier right tail whereas the negative/none effect has no right tails at all. **This indicates the treatment could be more effective on patients that have nasal mucus viscosity higher than usual (>= 2.5).**



Similarly, to demonstrate the treatment effect depends on the paper tissues. We will compare the proportion of patients who have positive effects and negative effects based on tissue use among active treatment groups. **From the graph below, we cannot see a clear indication of the relationship between the use of tissues and treatment effect.**



However, we could confirm that the most medically serious cases are those patients who buys a large amount of paper tissues by visualizing the relationships between nosebleeds and nosebleeds last year with tissue.use



Hospitalization for nosebleed may depend on local medical practice. Does this have any impact? How can you understand this?

Intuitively whether nosebleeds require hospitalization heavily depends on local medical practices. For example, in humid area, serious nosebleeds could be extremely rare which requires hospitalization. However, in dry areas, the nosebleeds at the same level may not be considered hospitalization at all. Both of the decisions are made relies on local medical practice. And there could be much more factors than the humid level. Those biases will definitely have some impact on how we measure the performance of the treatments. For example, the patients who live in dry areas tend to have more nosebleeds than the patients who live in humid areas.

How might you predict the rate of nosebleed from the data that you have? What might a statistical model for this look like?

There could be multiple ways we could predict the rate of nosebleeds. From the simplest model, we could use Linear Regression or Non-Linear Regression to some complex model like Regression tree.

However, due to the small amount of the observations. I would prefer to use simple models. However, based on the summary below we could see the simple linear regression model perform poorly on our dataset. If we diagnose the data, we could see most of the variables are not statistically significant. This is most likely because of our categorical features.

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  0  1  2  3  4
##          0 40 10  2  0  0
##          1 23  4  6  1  1
##          2  0  0  1  0  0
```

```
##          3  0  0  0  0  0
##          4  0  0  0  0  0
##
## Overall Statistics
##
##                Accuracy : 0.5114
##                  95% CI : (0.4025, 0.6195)
##     No Information Rate : 0.7159
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0466
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            0.6349  0.28571  0.11111  0.00000  0.00000
## Specificity            0.5200  0.58108  1.00000  1.00000  1.00000
## Pos Pred Value         0.7692  0.11429  1.00000      NaN      NaN
## Neg Pred Value         0.3611  0.81132  0.90805  0.98864  0.98864
## Prevalence             0.7159  0.15909  0.10227  0.01136  0.01136
## Detection Rate         0.4545  0.04545  0.01136  0.00000  0.00000
## Detection Prevalence   0.5909  0.39773  0.01136  0.00000  0.00000
## Balanced Accuracy      0.5775  0.43340  0.55556  0.50000  0.50000
```

```
##
## Call:
## lm(formula = nosebleeds ~ . - effect - subject - duration, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4630 -0.4603 -0.2265  0.0808  6.2774
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.27929    0.37360  -0.748 0.455239
## armPLACEBO        0.22358    0.10211   2.189 0.029243 *
## countryB         -0.52250    0.20961  -2.493 0.013153 *
## countryC         -0.54453    0.45601  -1.194 0.233261
## countryD         -0.81146    0.22131  -3.667 0.000285 ***
## countryE         -0.30441    0.22865  -1.331 0.183964
## countryF          0.10840    0.23394   0.463 0.643380
## countryG          0.11768    0.34255   0.344 0.731408
## countryH         -0.82119    0.20427  -4.020 7.17e-05 ***
## countryI         -0.61904    0.73331  -0.844 0.399170
## countryJ         -1.00232    0.69980  -1.432 0.152980
## eye.colourBLUE    0.66386    0.31534   2.105 0.036007 *
## eye.colourBROWN   0.84081    0.75359   1.116 0.265319
## tissue.useMEDIUM -0.13452    0.10508  -1.280 0.201355
## previous.year     0.23961    0.09528   2.515 0.012367 *
## mucus.viscosity   0.02791    0.04395   0.635 0.525881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9443 on 340 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1081
## F-statistic: 3.869 on 15 and 340 DF,  p-value: 2.098e-06
```

Instead, we could use a groupwise linear regression. **We first assign the samples into their subgroups based on categorical variables. Then using a regression model to make prediction of nosebleeds rate based on values of continuous variables only in the subgroups.** This could reduce the interaction effect between irrelevant features to minimum. For some of the NA coefficients in the following table, this is because the corresponding variable has the same value for all subgroup samples. Which could be resolved by increasing the sample size to include more various samples.

```
##        arm country eye.colour tissue.use (Intercept) previous.year
## 1: PLACEBO       I      BROWN       HIGH  18.1101190            NA
## 2:  ACTIVE       E       BLUE     MEDIUM  -0.9075028     0.4380829
## 3:  ACTIVE       J      BROWN     MEDIUM  -0.7511021     0.2946762
## 4: PLACEBO       J      BROWN     MEDIUM   0.8981954    -0.1280218
## 5:  ACTIVE       J      BROWN       HIGH   2.0457438    -0.4487576
## 6: PLACEBO       J      BROWN       HIGH   0.0000000            NA
##    mucus.viscosity
## 1:      -4.9717287
## 2:       0.0314280
## 3:       0.1559851
## 4:      -0.1033708
## 5:      -0.2981852
## 6:              NA
```

How can you use such a statistical model to simulate a Phase III trial? What inputs would it need, how would you generate them, and what outputs would it have?

To use this statistical model to simulate a Phase III trial. We need to generate the inputs first. The inputs we need would be the treatment groups, country, eye colour, tissues, nosebleeds in the last year and the mucus viscosity. For categorical variables, we could simply generate them using combinations. Then for the continuous data like mucus viscosity and nosebleeds in the previous year. We want to generate them from a normal distribution with mean and standard error to be the existing mean and standard error of the data subgroups (grouped by categorical variables).

For the outputs, we will get the predicted nosebleeds with unit times/year. Then with the predictive model and the generated data, we could simulate a Phase III trial and visualize the treatment effect. However, this method will have lots of disadvantages. It will reinforce the patterns of the observed data (overfit), where a real Phase III trial could bring more observations and include patterns we did not yet observe.