Statistiek 1

Rienk Heins

2/10/2020

Dietary effects on eosinophilic esophagitis in adults EDA

Goal

Eosinophilic esophagitis(EoE) is a chronic, allergic inflammatory disease of the esophagus. A study was performed to view the effects of dietary changes and elimination diets on EoE. Using the data from this study the following question will be tried to answer:

Can the allergic effect of eosinophilic esophagitis in a patient be predicted using machine learning with the dietary composition of foods and nutrients of this patient?

Obtaining data

The data for this EDA was obtained by a research done by Simone R.B.M. Eussen and Marleen T.J. Van Ampting from the Danone Nutrica Research in Utrecht, Willemijn E. de Rooj and Albert J. Bredenoord from the Amsterdam UMC, Sanne Wielders from the university on Wageningen and the Danone Nutrica Research in Utrecht and Berber Vlieg-Boerstra from the OVLG in Amsterdam and the Hanze University in Groningen. The title of their was: Dietary composition in adult eosinophilic esophagitis patients is related to disease severity. The goal of the research was to find the immunomodulatory role of nutrients and foods in adult eosinophilic esophagitis(EoE) before and during elimination diets.

A call was made with Berber Vlieg-Boerstra wherin she explained the data further and informed that the study wasn't fully completed yet. Therefore some of the data is not available at the time this EDA was made(last updated 20/09/2020), so some columns will be deletes purely because they are not available yet.

First the data has to be loaded in. The haven library is used to read in de .sav file.

```r
library(haven)
raw_Data <- read_sav("Datafile.sav")
```

The standard data contains 879 columns, so it isn't loaded into the pdf file as it would take up way to much space.

Cleaning of the data

After visualizing the data it is seen that there is no information about the patients after patient number 52. Since all the data after this patient is NA they will be removed since they don't give any information.

Also there is no data of patient number 21 and some patients dropped out during the study as there is information of them at the start, but no information after six weeks. These patients will also be thrown out of the data set.

Further a lot of the 879 columns contain NA values since the study hasn't been fully completed yet. These columns will therefore not be used in this project. Also the values of nutrients have already been set to value per 1000 kCal, thus these normalized values will be used instead of the normal values, cutting the columns containing these normal values in the process.

```r
# removing every record after patient number 52(SET_IDnr == 52), which is the 40th entry
# removing patients that dropped out
no.data.patients <- c(10, 14, 23, 30, 32, 41:79)
# selecting columns with information about the patients
patient.characteristics <- c(1, 3, 5, 6, 64)
# variables about primary outcome and diet of patient
data.variables <- c(630, 631, 737:853)
Data <- raw_Data[-no.data.patients, c(patient.characteristics, data.variables)]
# Changing gender to man and women as the raw data had gender as 0 and 1
Data$Gender <- factor(Data$Gender, labels = c("Women", "Man"))
```

The data contains columns with the peak allergic reaction at the start of the study(baseline) and after six weeks, but there are no columns containing the change in allergic reaction. Thus this column has to be made as this is the main variable to consider.

```r
# Calculating the difference in peakEos
peakEosDiff <- raw_Data$PeakEosSixwk_Max[-no.data.patients] - raw_Data$PeakEosBaseline_Max[-no.data.pati
# Normalizing the data
LNpeakEosDiff <- scale(peakEosDiff, 2)[,1]
# Adding data to data
EosDiffData <- data.frame("peakEosDiff" = peakEosDiff,
                          "LNpeakEosDiff" = LNpeakEosDiff)
cleanData <- cbind(Data, EosDiffData)
```

visualization

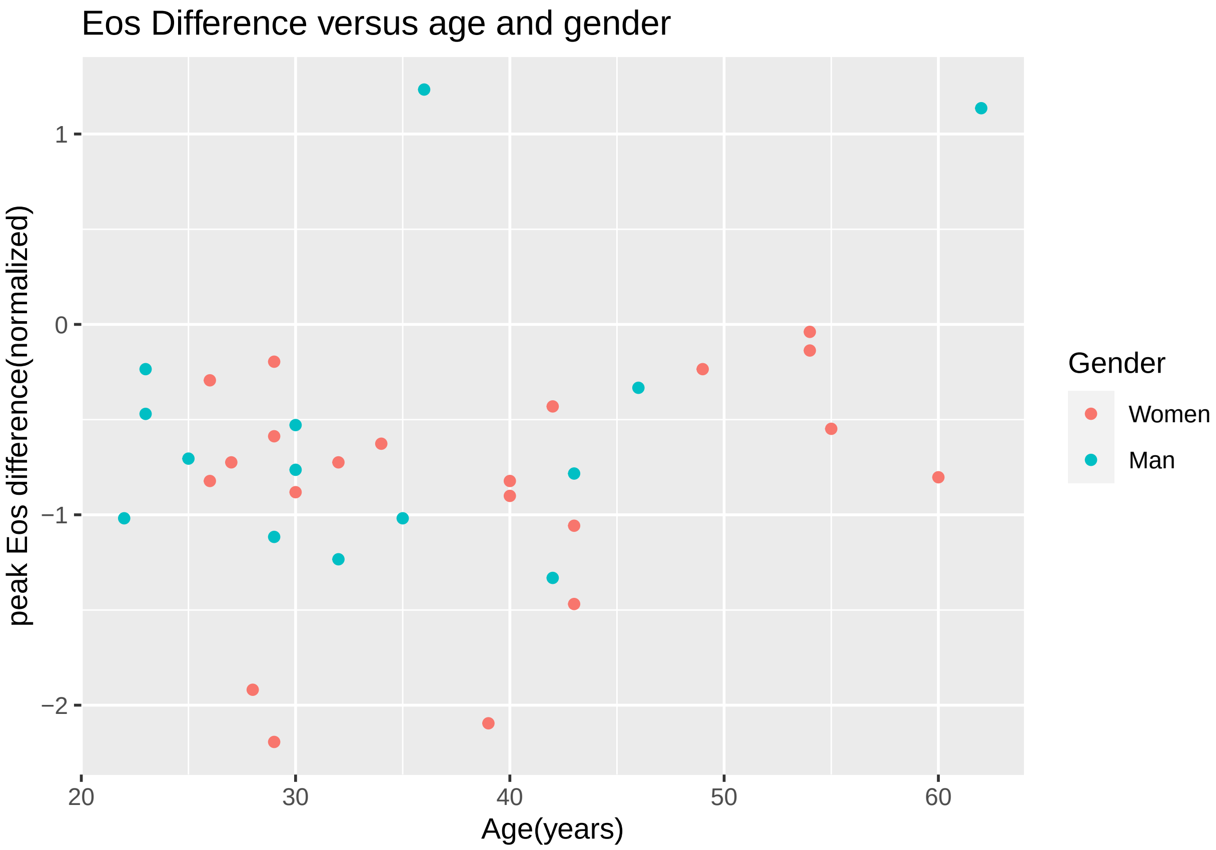
Before visualization it has to be determined which variables correlate with the primary variable. This will give a representation of variables which are worth looking into further as they seem to affect the main variable of the allergic reaction.

```r
correlation <- cor(x = cleanData$LNpeakEosDiff, y = Data[-3], use = "complete.obs")
sort(correlation[1,])
```

```
##             Ekcal.sixwk.VitC       LN_PeakEosBaseline_Max
##                 -0.564078063                 -0.555626952
##                  Dif_Zinc_mg     Ekcal.B.Carbohydrates.gr
##                 -0.460919628                 -0.427750612
## Ekcal.sixwk.Carbohydrates.gr              Dif_Calcium_mg
##                 -0.425926269                 -0.388637744
##              Ekcal.B.VitC.gr      Ekcal.sixwk.Folateequiv
##                 -0.365425227                 -0.364151560
##   Ekcal.sixwk.Monodisacch.gr           Ekcal.B.Retinol.gr
##                 -0.361486367                 -0.350495985
##             Dif_Irontotal_mg                Dif_Potassium
##                 -0.343482871                 -0.341470948
##             Dif_Potassium_mg           Dif_Folateequiv_mg
##                 -0.341470948                 -0.316099153
##        Ekcal.B.Nicotinacid.gr          Ekcal.sixwk.Folate
##                 -0.310057703                 -0.306202598
##                Dif_Sat.Fat_g         Ekcal.B.monodisacch
##                 -0.302754313                 -0.299109252
##          Ekcal.sixwk.Ironheam               Dif_Folate_mg
##                 -0.290922723                 -0.278518962
##             Ekcal.B.Sodium.gr        Ekcal.sixwk.Irontotal
##                 -0.270577534                 -0.247745607
##        Ekcal.sixwk.Potassium             Ekcal.sixwk.Zinc
##                 -0.242521041                 -0.232581482
##                  Dif_VitC_mg                 B.Alcohol.gr
##                 -0.220238020                 -0.211389345
##        Ekcal.sixwk.Folicacid           Dif_Ironnonheam_mg
##                 -0.200536381                 -0.192662817
##      Ekcal.sixwk.Ironnonheam             Dif_Folicacid_mg
##                 -0.192018692                 -0.189088091
##               Dif_Protein_gr                Dif_Iodine_mg
##                 -0.187812082                 -0.187210087
##      Ekcal.sixwk.Nicotinacid                 Dif_VitB6_mg
##                 -0.181236171                 -0.180011356
##       Ekcal.sixwk.Phosphorus            Dif_Phosphorus_mg
##                 -0.172968447                 -0.167406334
##           Ekcal.B.Alcohol.gr             Dif_Polysacch_gr
##                 -0.166955578                 -0.166767734
##           Ekcal.sixwk.Iodine    Ekcal.sixwk.Polysaccch.gr
##                 -0.165806758                 -0.152786728
##              Dif_Selenium_mg          Dif_Carbohyrdate_gr
##                 -0.152388529                 -0.146377355
##                Dif_Copper_mg               Ekcal.B.RAE.gr
##                 -0.129534462                 -0.128506973
##         Ekcal.sixwk.Selenium          Ekcal.B.Ironheam.gr
##                 -0.128151186                 -0.124116614
##            Ekcal.sixwk.VitB6              Ekcal.sixwk.DHA
##                 -0.121754750                 -0.118768896
##              Ekcal.sixwk.EPA              Dif_Ironheam_mg
##                 -0.097590608                 -0.096075299
##                 Dif_VitB1_mg      Ekcal.B.Dietaryfiber.gr
##                 -0.095242574                 -0.093488611
##        Ekcal.B.Polysaccch.gr                      Neocate
##                 -0.086012072                 -0.062920019
##                 Dif_VitB2_mg                  Ekcal.B.EPA
##                 -0.046079511                 -0.034247677
##                  Dif_VitE_mg       Ekcal.sixwk.Sat.fat.gr
##                 -0.028413501                 -0.026153301
##                  Ekcal.B.DHA        Ekcal.B.phosphorus.gr
##                 -0.015069097                 -0.008280975
##      Ekcal.B.Linoleicacid.gr             Dif_Magnesium_mg
##                 -0.006739630                 -0.003513541
##                  Ekcal.B.ALA                Dif_VitB12_mg
##                  0.010533136                  0.015034376
##          Ekcal.sixwk.Retinol             Ekcal.B.PUFAS.gr
##                  0.015287067                  0.015587182
##            Ekcal.B.Folate.gr         Ekcal.B.Irontotal.gr
##                  0.020147723                  0.022615141
##            Ekcal.B.VitB12.gr          Ekcal.B.Selenium.gr
##                  0.034690678                  0.039131065
##         Ekcal.B.Magnesium.gr           Ekcal.sixwk.VitB12
##                  0.042219043                  0.050401589
##       Ekcal.B.Ironnonheam.gr           Ekcal.sixwk.Sodium
##                  0.053986763                  0.057407619
##            Ekcal.B.Iodine.gr             Ekcal.B.VitB2.gr
##                  0.058433273                  0.067973702
##           Dif_Nicotinacid_mg       Ekcal.sixwk.Alcohol.gr
##                  0.071755793                  0.072508701
##      Ekcal.B.Retinolequiv.gr       Ekcal.sixwk.Protein.gr
##                  0.076317916                  0.082295529
##         Ekcal.B.Folicacid.gr        Ekcal.sixwk.Magnesium
##                  0.100870652                  0.106613630
##       Ekcal.B.Folateequiv.gr             Ekcal.B.VitB6.gr
##                  0.109387377                  0.111984170
##              Ekcal.B.VitD.gr                   Dif_ALA_mg
##                  0.121220535                  0.125983489
##            Ekcal.sixwk.VitB1                    Dif_Fat_g
##                  0.126963117                  0.131489705
##             Ekcal.B.VitB1.gr         Ekcal.B.Potassium.gr
##                  0.157272498                  0.165232411
##              Ekcal.sixwk.ALA  Ekcal.sixwk.Linoleicacid.gr
##                  0.181830285                  0.188064418
##  Ekcal.sixwk.Dietaryfiber.gr               Dif_Alcohol_gr
##                  0.217573311                  0.227074911
##                     SET_IDnr              Ekcal.B.VitE.gr
##                  0.238094207                  0.239530971
##             Ekcal.sixwk.VitE           Ekcal.sixwk.Copper
##                  0.250479796                  0.252333674
##            Ekcal.sixwk.VitB2          Dif_Linoleicacid_mg
##                  0.255151362                  0.256247424
##              Ekcal.B.Zinc.gr          Dif_Dietaryfiber_gr
##                  0.265651709                  0.270793064
##         Ekcal.sixwk.PUFAS.gr                  Dif_VitD_mg
##                  0.291263097                  0.297488005
##                Dif_Sodium_mg                 Dif_PUFAS_mg
##                  0.304839496                  0.333634864
##            Ekcal.B.Copper.gr              Ekcal.B.Protein
##                  0.338829374                  0.350538092
##                          Age                   Dif_EPA_mg
##                  0.358758708                  0.359124964
##                   Dif_DHA_mg           Ekcal.sixwk.Fat.gr
##                  0.360839952                  0.414590365
##               Ekcal.B.Fat.gr              Ekcal.sixwk.RAE
##                  0.419020824                  0.420435084
##          Ekcal.sixwk.Calcium         Dif_Retinolequiv_mcg
##                  0.428647012                  0.530598231
##           Ekcal.B.Sat.fat.gr              Dif_Retinol_mcg
##                  0.539686361                  0.540683969
##     Ekcal.sixwk.Retinolequiv             Ekcal.sixwk.VitD
##                  0.544396685                  0.561761171
##             Ekcal.Calcium.gr                   Dif_RAE_mg
##                  0.571711739                  0.591283264
##          LN_PeakEosSixwk_Max
##                  0.646390028
```

Age and gender might have an effect on the allergic reaction of the patient, so these variables will be plotted out.

```
library(ggplot2)
ggplot(data = cleanData) + geom_point(aes(Age, LNpeakEosDiff, color = Gender)) + ggtitle("Eos Differenc
```



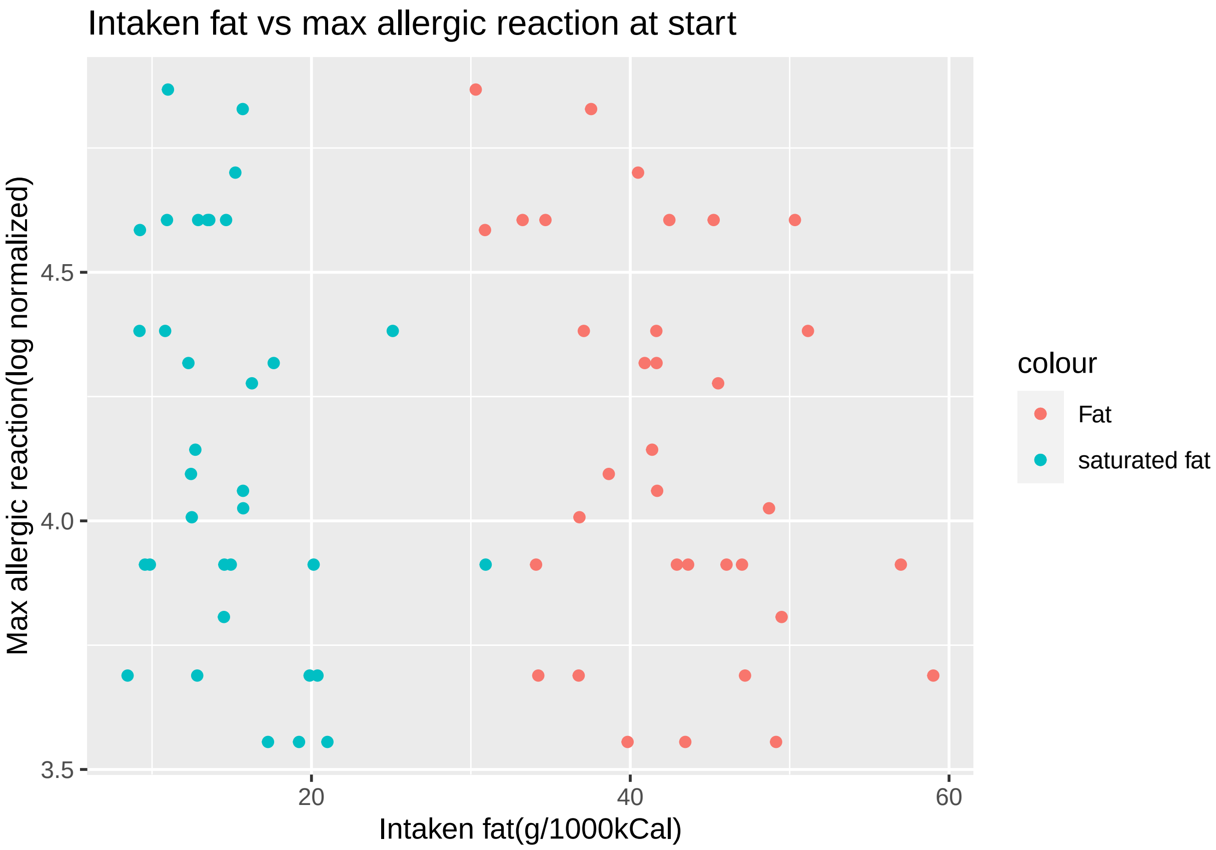
As seen in the figure above, age might have a slight impact on the difference in reaction as after an age of 45 there aren't anymore decreases higher than 1. So at a higher age the dietary changes might not have as much effect as it would have on a younger age. Furthermore gender doesn't really seem to have an effect as the outer limits are specific women for decreases and men for increases, but these seem to be outliers as the rest of the patients seem to stay around the same level.

Another variable to look at are saturated fat and fat in general as they give some of the higher levels in the correlation test(0,54 B.sat.fat, 0,42 and 0,41 for fat on both time periods). Also saturated fat is known as a nutrients that's bad for health when eaten in bigger quantities. In the following graph the max allergic reaction will be set against the amounts of fat in the diet.

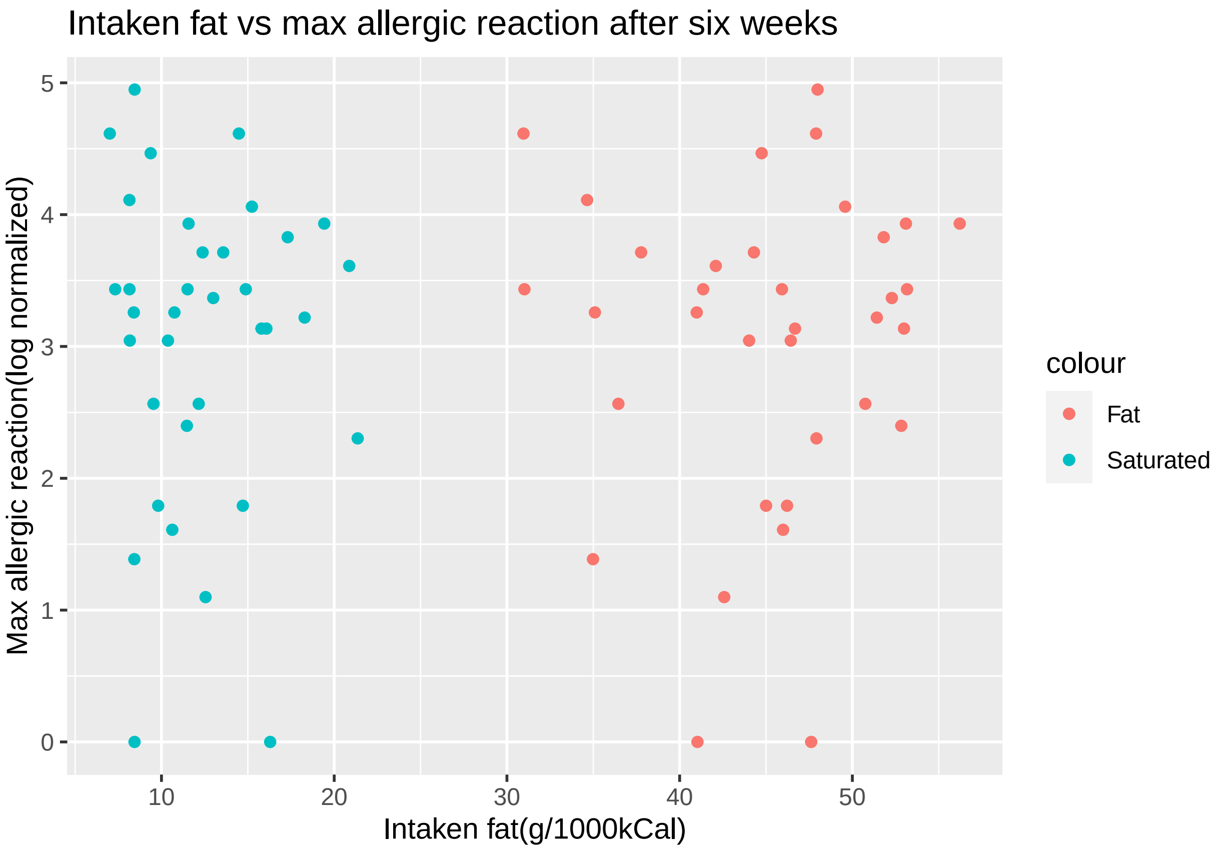
```r
fatBaseline <- geom_point(aes(Ekcal.B.Fat.gr, LN_PeakEosBaseline_Max, color = "Fat"))
fatSixwk <- geom_point(aes(Ekcal.sixwk.Fat.gr, LN_PeakEosSixwk_Max, color = "Fat"))
satFatBaseline <- geom_point(aes(Ekcal.B.Sat.fat.gr, LN_PeakEosBaseline_Max, color = "saturated fat"))
satFatSixwk <- geom_point(aes(Ekcal.sixwk.Sat.fat.gr, LN_PeakEosSixwk_Max, color = "Saturated"))
ggplot(data = cleanData) + fatBaseline + satFatBaseline + ggtitle("Intaken fat vs max allergic reaction
```

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```



```
ggplot(data = cleanData) + fatSixwk + satFatSixwk + ggtitle("Intaken fat vs max allergic reaction after
```



Something to look at in the figure above is that the highest peaks seem to fall under the patients with a lower fat intake, mostly under 15 grams/1000 kCal for saturated fat and 50 grams/1000 kCal for total fat. As there are almost no peaks for saturated fat above 4.5 at baseline, and no peaks above 4 after six weeks with this minimum. Same for total fat where there is only one peak above 4.5 for 50 grams in the baseline and zero after six weeks with a peak above 4. One thing to note here is that at baseline there are not a lot of points in general above these minimum values.

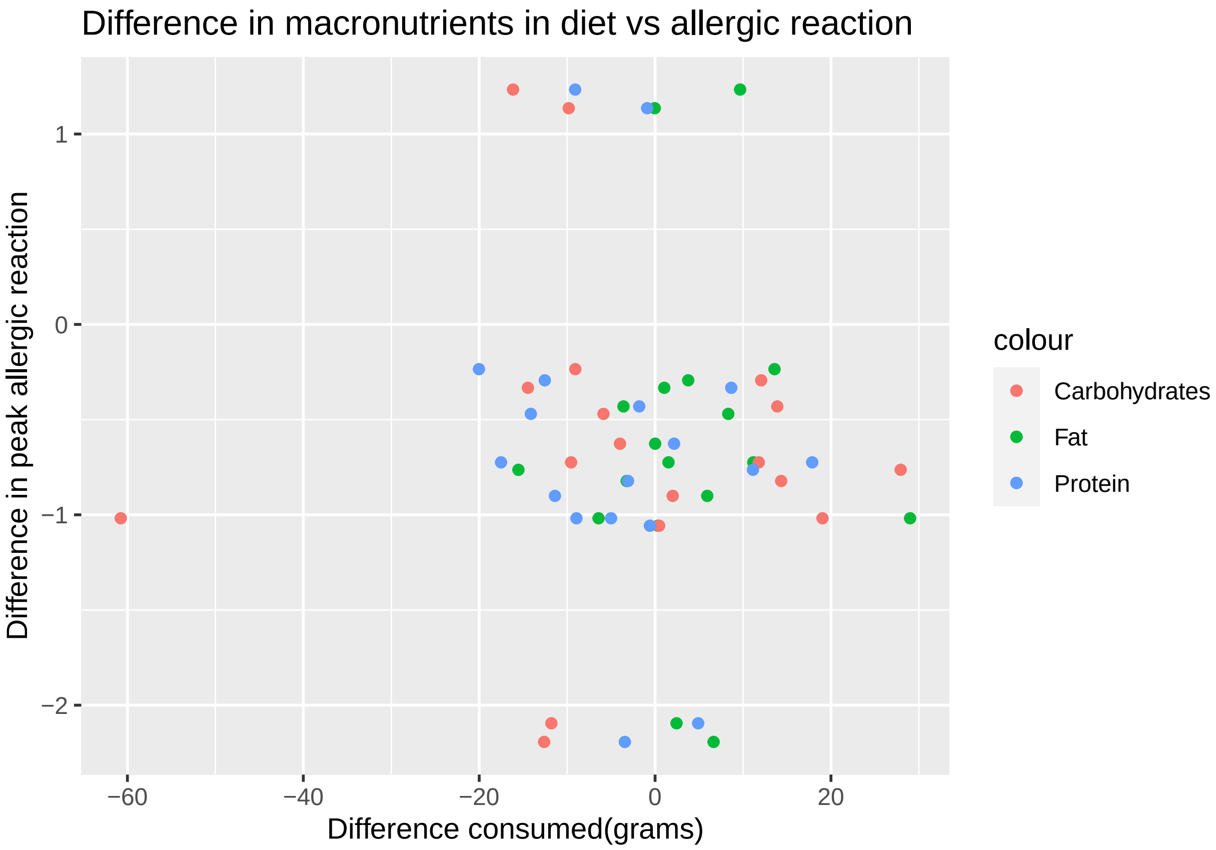
Something to look at as well is the difference in macro nutrients, these being the total protein, fat and carbohydrates as these by far make up the most pure mass in grams of a diet and serve as the sources of energy and building blocks for the body.

```
fatdiffpoints <- geom_point(aes(Dif_Fat_g, LNpeakEosDiff, color = "Fat"))
carbdiffpoints <- geom_point(aes(Dif_Carbohyrdate_gr, LNpeakEosDiff, color = "Carbohydrates"))
proteindiffpoints <- geom_point(aes(Dif_Protein_gr, LNpeakEosDiff, color = "Protein"))
ggplot(data = cleanData) + fatdiffpoints + carbdiffpoints + proteindiffpoints + ggtitle("Difference in
```

```
## Warning: Removed 17 rows containing missing values (geom_point).

## Warning: Removed 17 rows containing missing values (geom_point).

## Warning: Removed 17 rows containing missing values (geom_point).
```



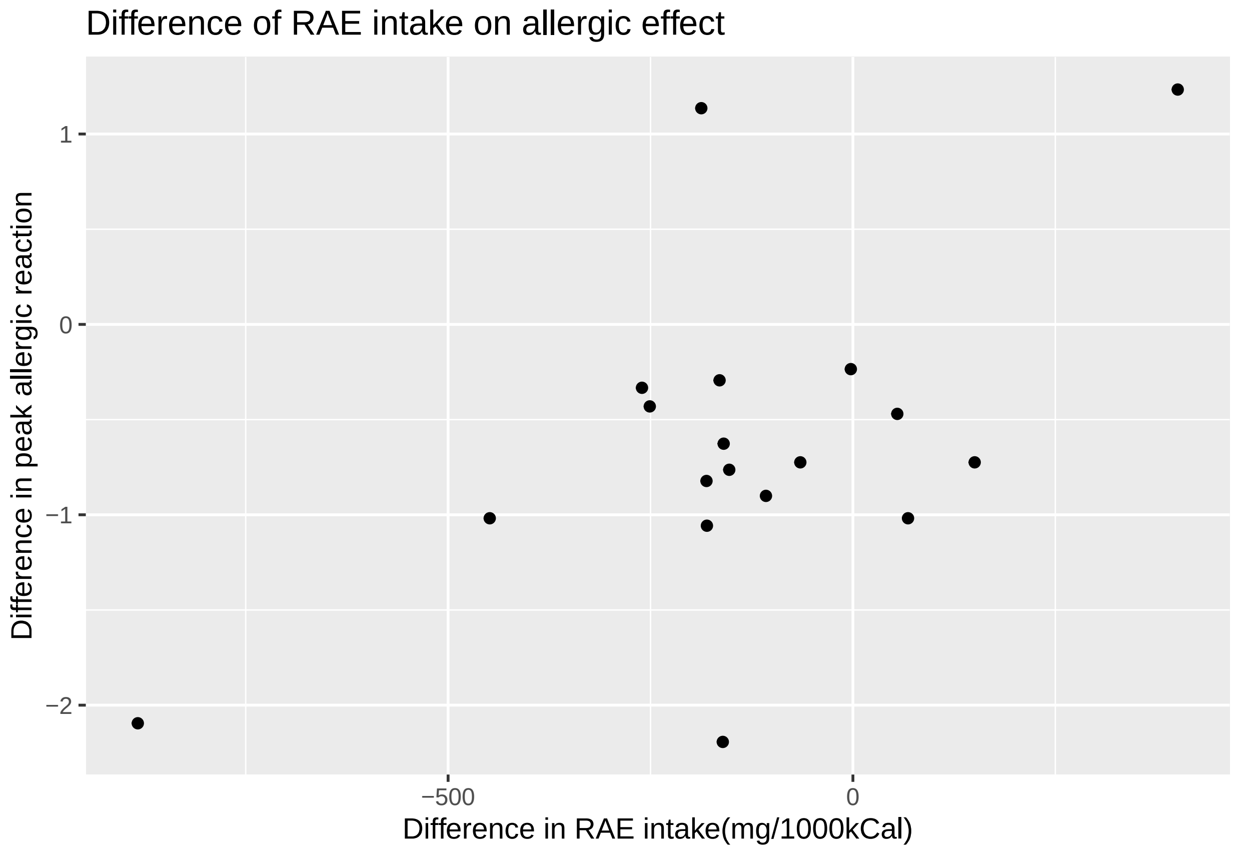
Even though there doesn't seem to be a clear distinction in the data through macro nutrients there is a slight correlation as above a peak allergic reaction decrease of 0,5 most fall under a decrease in protein intake and an increase of carbohydrates intake. Even though this doesn't mean there is a big impact as for the rest of the points they seem all pretty centered together this might suggest that a lower protein intake and higher carbohydrate intake can have an effect on the allergic reactions of a patient.

For the following plots there will be looked at the difference in intake for some minerals and vitamins that seem to have a high correlation with the difference in allergic reaction according to the values from the cor() function.

One of these variables is retinol activity equivalents(RAE), which accounts for the different bioactivities of retinol and provitamin A carotenoids, which all eventually will be converted to retinol.

```
ggplot(data = cleanData) + geom_point(aes(Dif_RAE_mg, LNpeakEosDiff)) + ggtitle("Difference of RAE intak
```

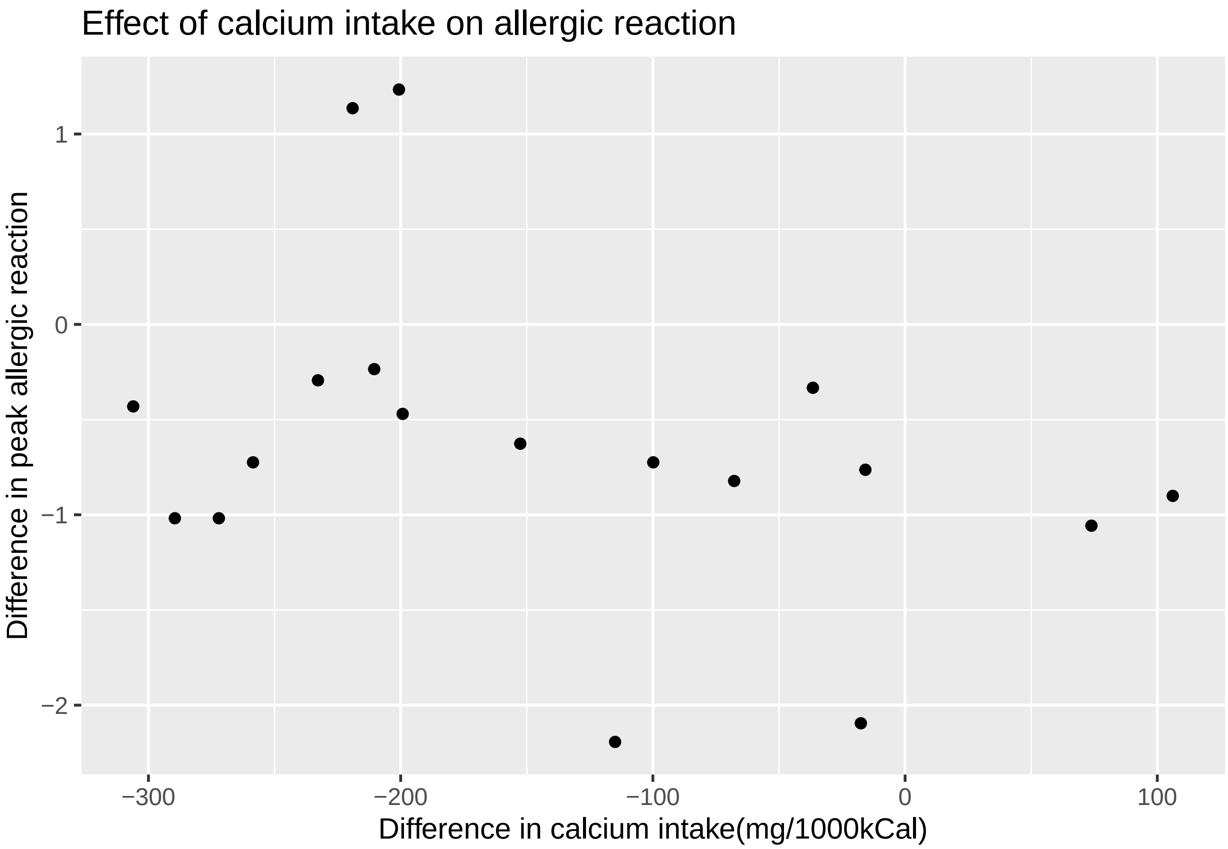
```
## Warning: Removed 17 rows containing missing values (geom_point).
```



One of the minerals that shows correlation and was found as a possibly important variable by the original study was calcium.

```
ggplot(data = cleanData) + geom_point(aes(Dif_Calcium_mg, LNpeakEosDiff)) + ggtitle("Effect of calcium 
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```

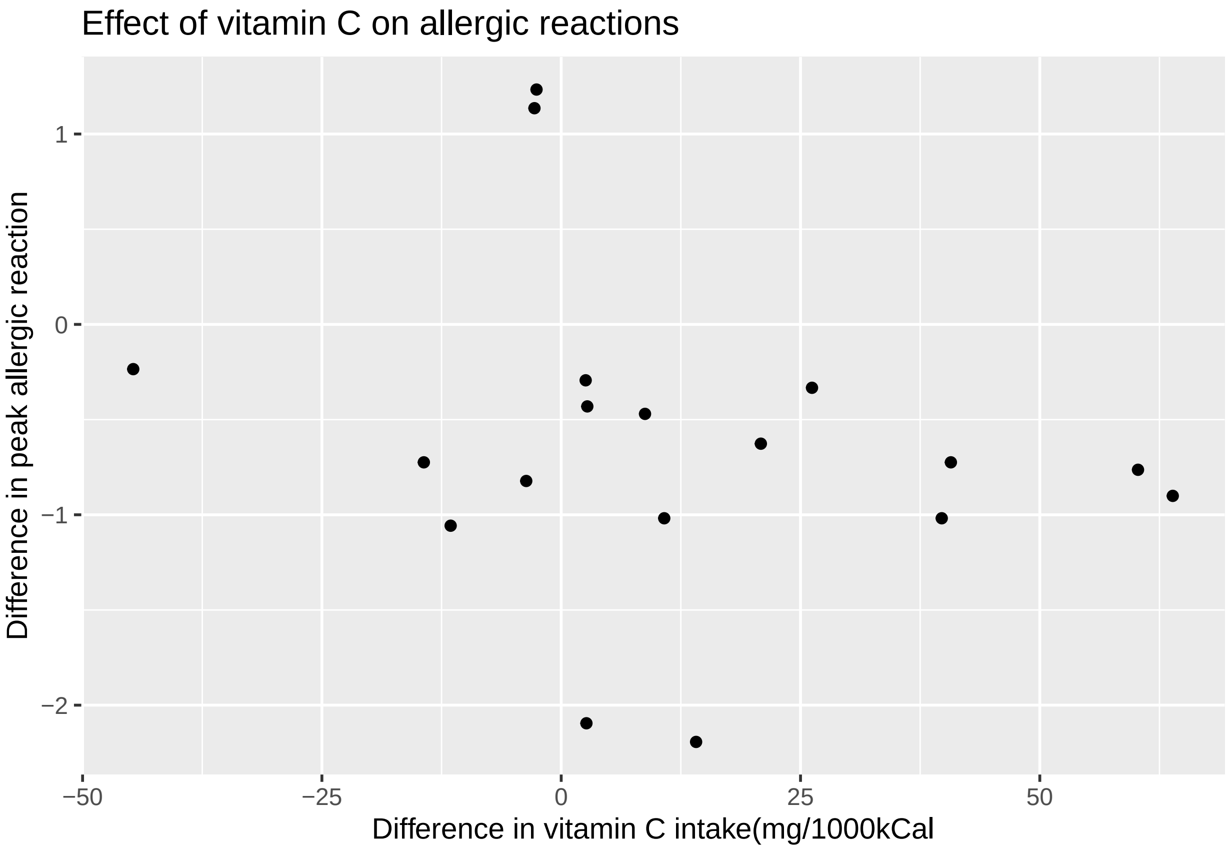


The lowering of calcium intake seems to slowly decrease the lowering of allergic reaction till a decrease of 250 mg/1000 kCal, an increase in calcium seems to decrease the allergic reaction highly, but there are only 2 patients with a higher intake of calcium so this might also be pure coincidence.

Another vitamin to look at is vitamin C, as it shows correlation and is known to boost the immune system. Thus it might impact the response of allergic reactions.

```
ggplot(data = cleanData) + geom_point(aes(Dif_VitC_mg, LNpeakEosDiff)) + ggtitle("Effect of vitamin C on
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```

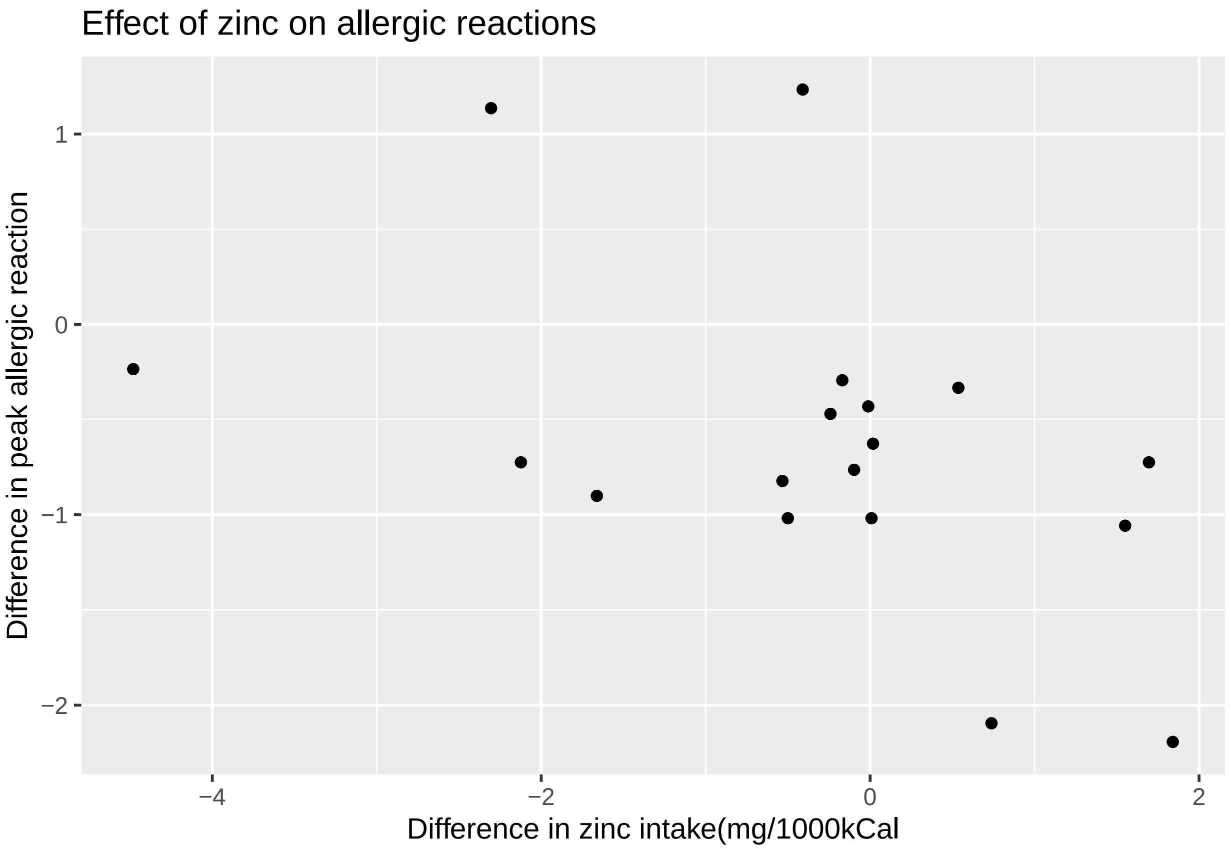


The difference in vitamin C intake seems to give a similar effect at most values as there are decreases of 0,5 and higher in allergic reaction at vitamin C drops of around 15 mg/1000 kCal to increases up to 65 mg/1000 kCal.

The last mineral to look at is zinc, as it also gives high correlation to the difference in allergic reaction.

```
ggplot(data = cleanData) + geom_point(aes(Dif_Zinc_mg, LNpeakEosDiff)) + ggtitle("Effect of zinc on all
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```



An increase in zinc seems to give a decrease in allergic reactions as most of the greater decreases come with a increase of zinc in the diet. These points also lay at a zinc decrease but at higher increases the allergic reaction decrease gets lower. Yet there aren't a lot of points in the areas of zinc increase en large zinc decreases so this isn't heavily supported.

Conclusions

As most correlated values don't really give clear values of how they impact the allergic reaction the effect probably is impacted by multiple values and their combinations. But higher carbohydrate, lower protein and higher zinc intake with minimum values of 50 g/1000 kCal of total fat and a 15 g/1000 kCal saturated fat seem to decrease allergic effects.