What urinary biomarkers in combination with PanRISC are most accurate at predicting pancreatic cancer

Nils Mooldijk

9/15/2021

What urinary biomarkers in combination with machine learning are most accurate at predicting pancreatic cancer?

Pancreatic ductal adenocarcinoma (PDAC) is an extremely lethal form of cancer. Less than 9% of diagnosed patients survive longer than 5 years. Because PDAC does not show symptoms in it's early stages, PDAC is mostly diagnosed too late, when the disease is already advanced to a locally advanced or metastatic disease. The big problem is that there are no useful bio-markers for detection in the early stages, when surgery is most effective. Though S. Debernardi et al. did establish a panel of bio-markers (LYVE1, REG1A, REG1B, and TFF1) that show promise for early detection of PDAC in urine (1).

Basic data of the patients

Some of the basic information can tell a lot about the entries in the data set. Perhaps there is a pattern or correlation between simple factors like age and gender. What is the age distribution of the entries? How is the division of confirmed cases between men and women? And does age play in role in that?

```
#Create a dataframe of the existing matrix so the rows and columns are no longer fixed.
DF_measured_data <- data.frame(measured_data)
```

Taking a look at the general data first:

```
summary(DF_measured_data)
```

```
##    sample_id   patient_cohort sample_origin      age        sex
##  S1     :  1   Cohort1:332    BPTB:409      Min.   :26.00   F:299
##  S10    :  1   Cohort2:258    ESP : 29      1st Qu.:50.00   M:291
##  S100   :  1                  LIV :132      Median :60.00
##  S101   :  1                  UCL : 20      Mean   :59.08
##  S102   :  1                                3rd Qu.:69.00
##  S103   :  1                                Max.   :89.00
##  (Other):584
##    diagnosis         stage                        benign_sample_diagnosis
##  Min.   :1.000          :391                                    :382
##  1st Qu.:1.000   III    : 76   Pancreatitis                     : 41
##  Median :2.000   IIB    : 68   Pancreatitis (Chronic)           : 35
##  Mean   :2.027   IV     : 21   Gallstones                       : 21
##  3rd Qu.:3.000   IB     : 12   Pancreatitis (Alcohol-Chronic)   : 11
##  Max.   :3.000   IIA    : 11   Cholecystitis                    :  9
##                  (Other): 11   (Other)                          : 91
##  plasma_CA19_9       creatinine          LYVE1
##  Min.   :    0.0   Min.   :0.05655   Min.   : 0.000129
##  1st Qu.:    8.0   1st Qu.:0.37323   1st Qu.: 0.167179
##  Median :   26.5   Median :0.72384   Median : 1.649862
##  Mean   :  654.0   Mean   :0.85538   Mean   : 3.063530
```

```
##  3rd Qu.:  294.0   3rd Qu.:1.13948   3rd Qu.: 5.205037
##  Max.   :31000.0   Max.   :4.11684   Max.   :23.890323
##  NA's   :240
##      REG1B               TFF1              REG1A
##  Min.   :   0.0011   Min.   :    0.005   Min.   :    0.00
##  1st Qu.:  10.7572   1st Qu.:   43.961   1st Qu.:   80.69
##  Median :  34.3034   Median :  259.874   Median :  208.54
##  Mean   : 111.7741   Mean   :  597.869   Mean   :  735.28
##  3rd Qu.: 122.7410   3rd Qu.:  742.736   3rd Qu.:  649.00
##  Max.   :1403.8976   Max.   :13344.300   Max.   :13200.00
##                                          NA's   :284
```

```
#taking a look at the age distribution of the entries.
ggplot(data = DF_measured_data, mapping = aes(x=age)) +
  geom_histogram(color="black", fill="white", bins = 50) +
  geom_vline(aes(xintercept=mean(age)), color="blue", linetype="dashed", size=1) +
  labs(x = "Age in years", y = "Number of patients")
```

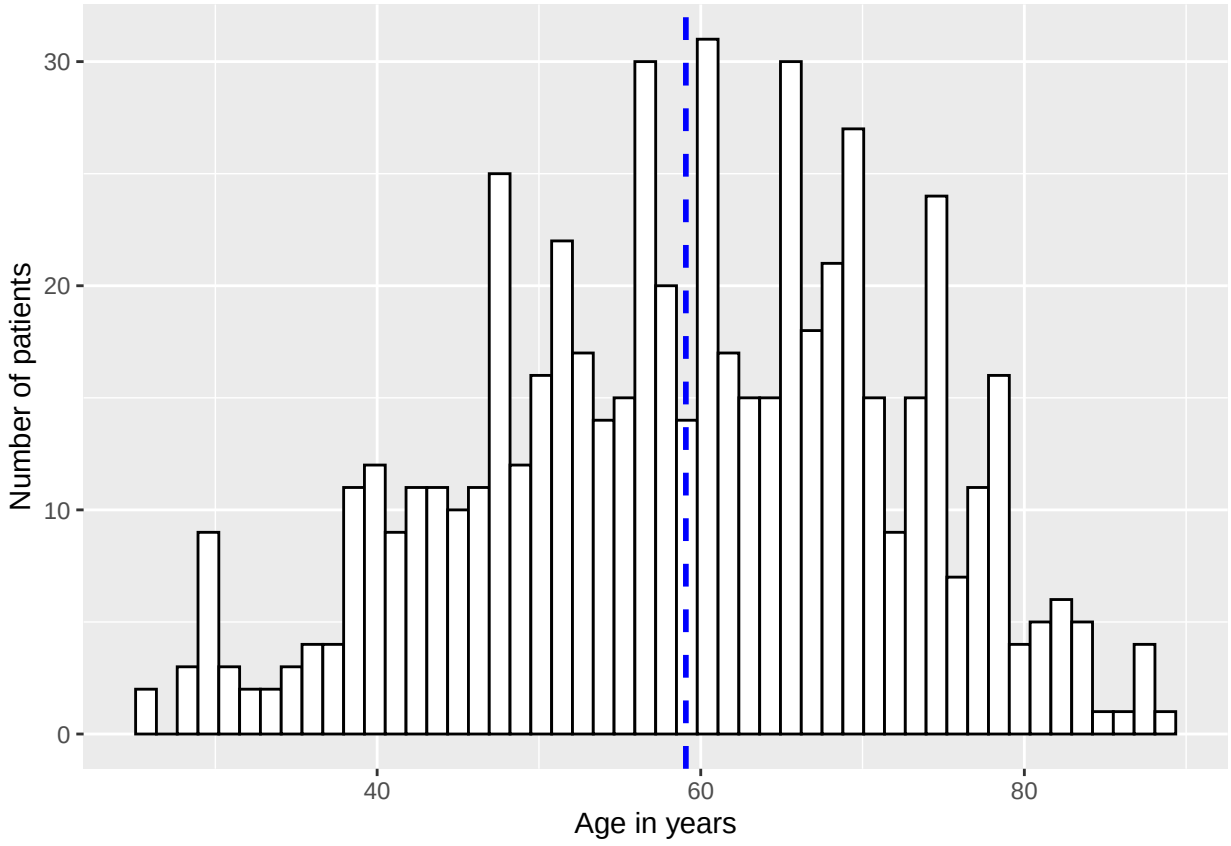

Figure 1 suggest that tested subjects are fairly evenly divided based on age. Though the distribution leans in favour of the elderly. Since the data contains both control, benign and PDAC groups, the above figure can be misleading. As it doesn't account for any group bias.

A dot plot, like figure 2 below, can be far more useful in separating the groups to find a pattern.

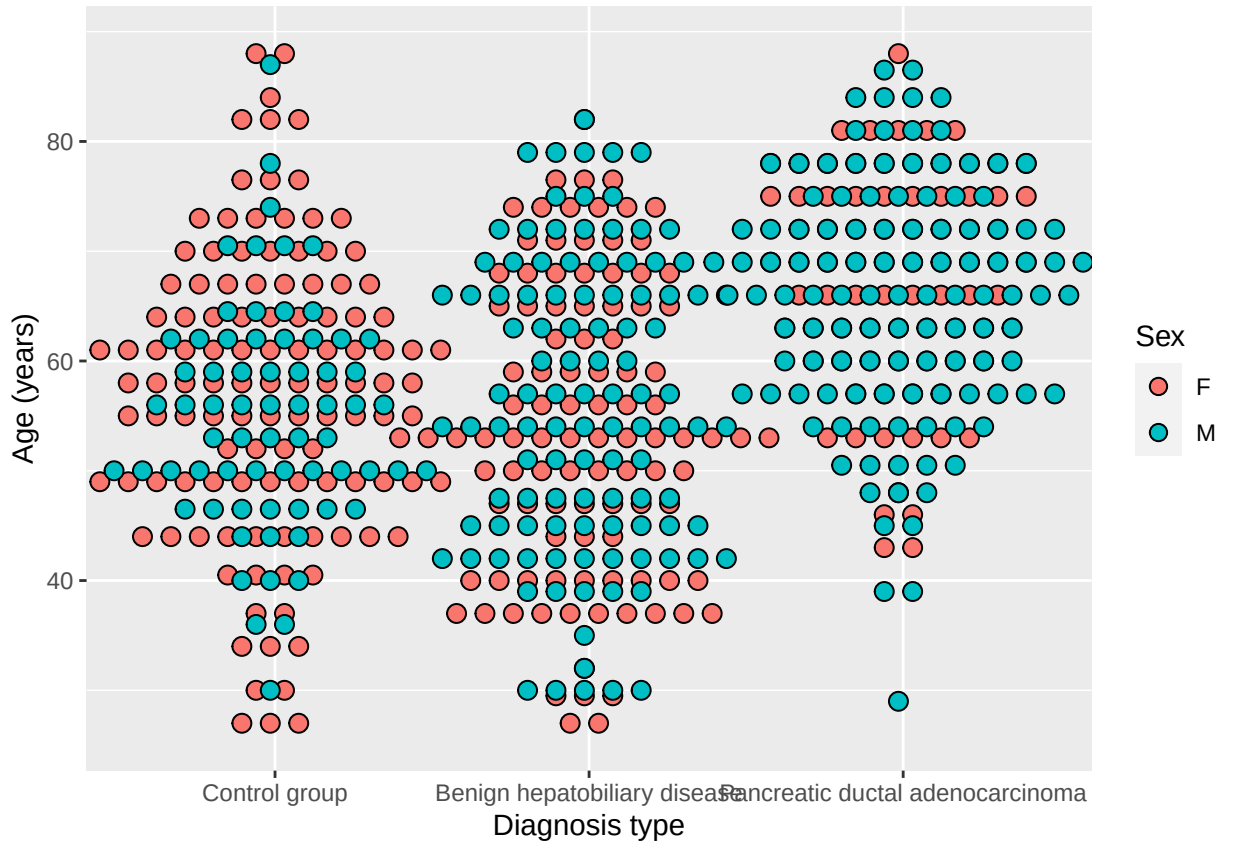

Figure 1: Patients grouped by diagnosis, distributed by age.

```
DF_measured_data$diagnosis <- as.factor(DF_measured_data$diagnosis)
ggplot(data = DF_measured_data, mapping = aes(x=diagnosis, y=age, fill=as.factor(sex))) +
          geom_dotplot(binaxis='y', stackdir='center', stackratio=1.5, dotsize=0.75, binwidth = 2.3) +
          scale_x_discrete(breaks = 1:3, labels=c("Control group","Benign hepatobiliary disease","Pancr
          labs(x = "Diagnosis type", y = "Age (years)", fill = "Sex" )
```

Based on the above plot (fig 2), there seems to be a trend for older men to be afflicted by PDAC. Both the control group and the benign group are more equally distributed according to age. Though this is a trend that doesn't necessarily mean that men are more susceptible to PDAC. Men tend to attend their general practitioner later in the course of a condition than women and this phenomenon is exacerbated by social class inequalities (2). The unequal representation of men in the PDAC group might be a consequence of men's unwillingness to attend their physician with early complaints, rather than them being more susceptible to PDAC. Because this factor is not relevant to which bio-marker is the clearest indicator of PDAC, the sex of the patient should be used with caution when feeding the data to a machine learning algorithm. As it could produce biased and therefore inaccurate results. Although this data cannot be set aside fully just yet. Further exploration into correlation between concentrations of certain bio-markers and sex is needed first.

Concentrations of biomarkers

There are several bio-markers found in urine that can indicate the body's function and state. To get an idea of which markers, if any, show promise in particular for indicating PDAC. Though some markers, like creatinine, are used to indicate kidney function rather than laying a direct link to PDAC, it's worth knowing if PDAC patients have a higher concentration of creatinine in their urine. Furthermore, it's important to establish the correlation between age and concentrations of bio-markers. With age comes more health risks.

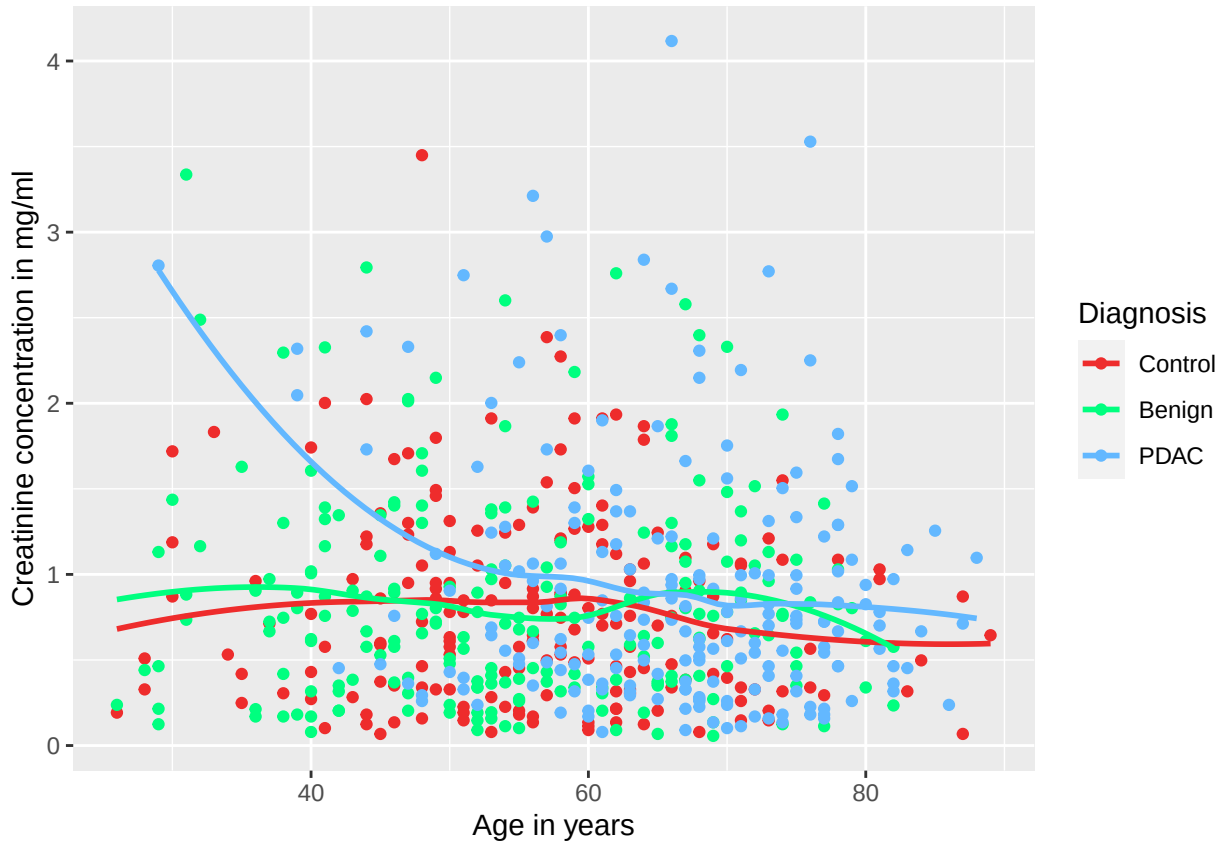

Figure 2: Correlation between age and concentration of creatinine.

```
ggplot(DF_measured_data, aes(age, creatinine)) +
  geom_point(aes(colour = factor(diagnosis))) +
  geom_smooth(se=F, aes(colour=factor(diagnosis))) +
  labs(y="Creatinine concentration in mg/ml",
       x="Age in years",
       color = "Diagnosis") +
  scale_color_manual(labels = c("Control", "Benign", "PDAC"), values = c("firebrick2", "springgreen", "
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

It seems that the creatinine concentration levels stay somewhat the same on average for all groups. The noticeable spike in the left hand side of the plot is thanks to the youngest PDAC patient having a high concentration of creatine. Though the concentration of creatinine in PDAC patients is slightly higher throughout, there is not such an obvious increase that could be a clear sign of PDAC.

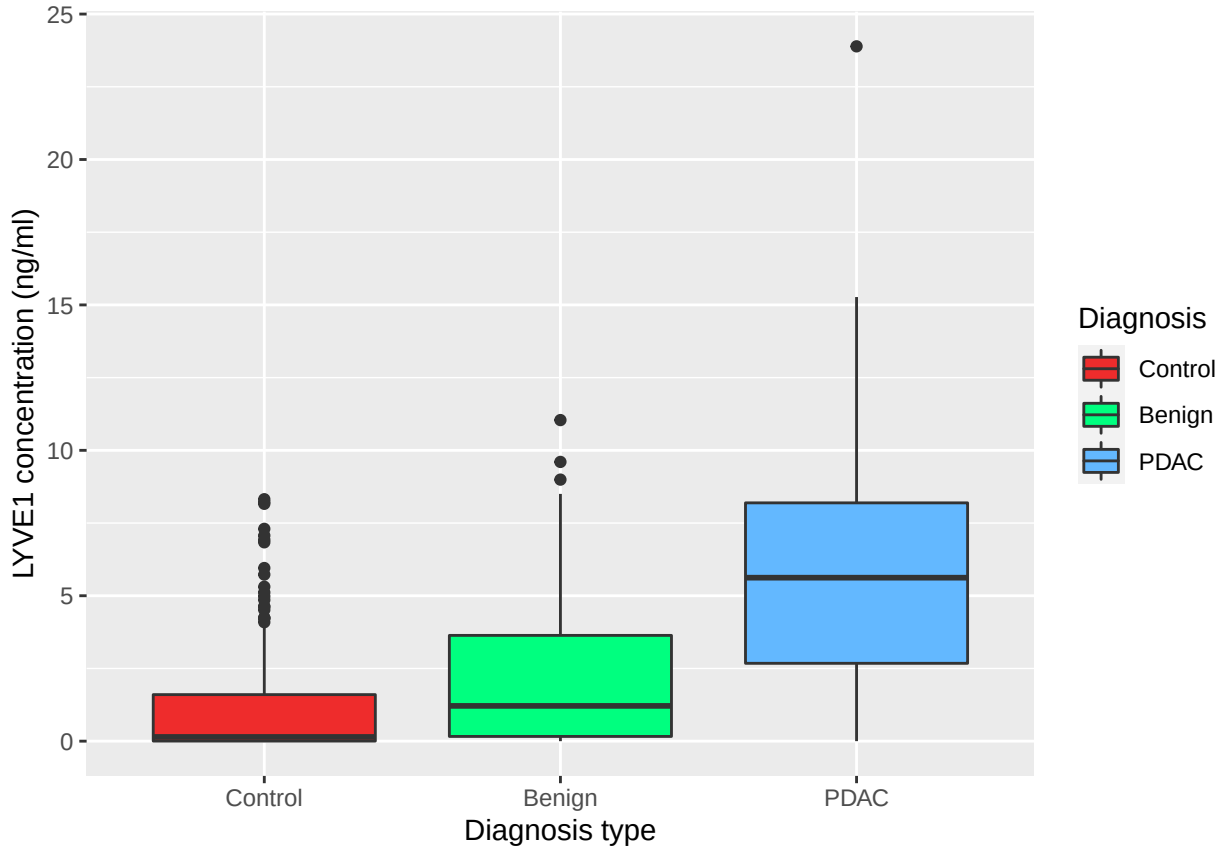

Figure 3: Concentration of LYVE1 in patients of different diagnosis.

Lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1) is a protein that may play a role in tumor metastasis. Therefore it has great potential as a marker for PDAC, and even other forms of cancer. Though the concentrations are sure to differ amongst the three main diagnosis groups.

```
ggplot(DF_measured_data, aes(x=diagnosis, y=LYVE1, fill=diagnosis)) +
  geom_boxplot() +
  scale_x_discrete(breaks = 1:3, labels=c("Control","Benign","PDAC")) +
  labs(x = "Diagnosis type", y = "LYVE1 concentration (ng/ml)", fill = "Diagnosis") +
  scale_fill_manual(values = c("firebrick2", "springgreen", "steelblue1"),
                    labels = c("Control","Benign","PDAC"))
```

Fig 4. shows an expected results for a bio-markers associated with tumor development. The control group has a low concentration, the benign group slightly higher, and the PDAC group has a very high concentration. Though there are a few outliers in the control group, the majority of the group seems to be between 0 and 2 ng/ml.

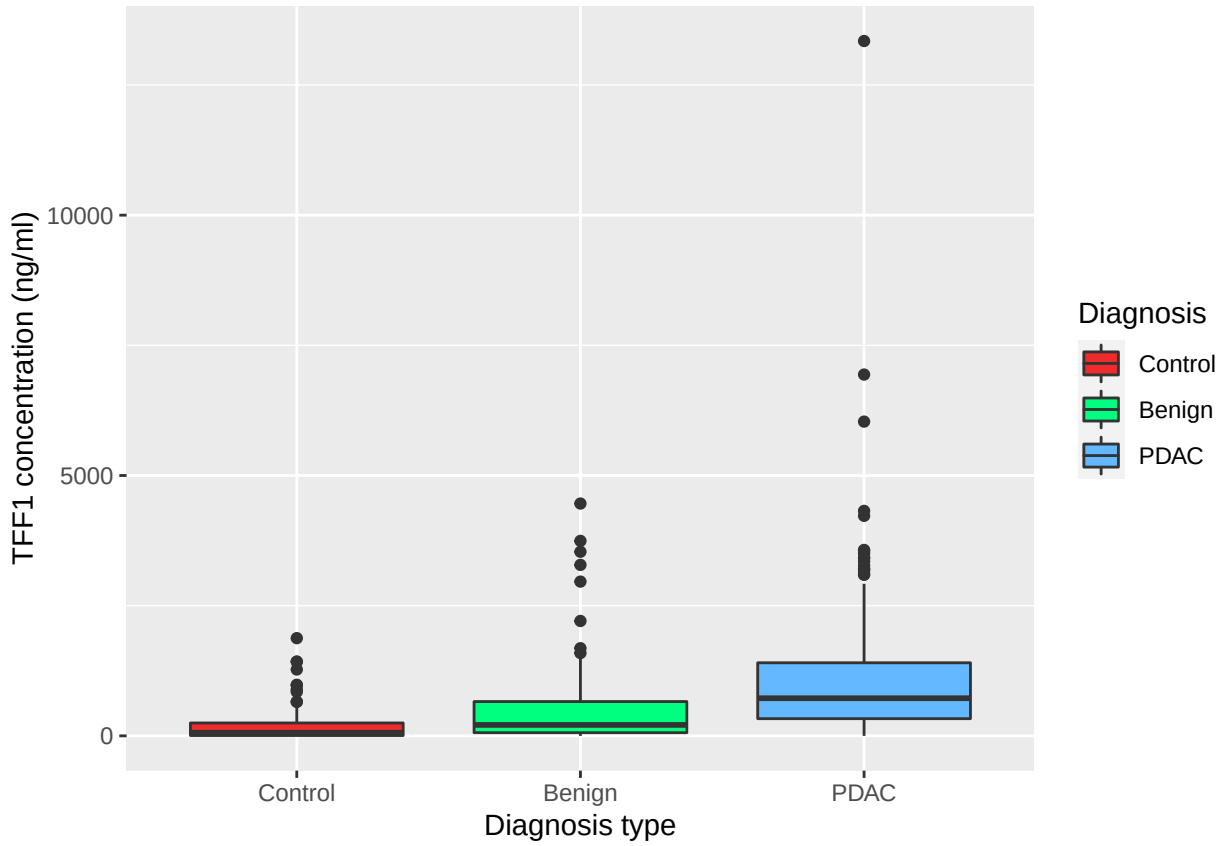

Figure 4: Concentration of TFF1 in patients of different diagnosis.

Urinary levels of Trefoil Factor 1 (TFF1) may be related to regeneration and repair of the urinary tract. Like with LYVE1, a higher concentration in patients with PDAC diagnosis is expected.

```
ggplot(DF_measured_data, aes(x=diagnosis, y=TFF1, fill=diagnosis)) +
  geom_boxplot() +
  scale_x_discrete(breaks = 1:3, labels=c("Control","Benign","PDAC")) +
  labs(x = "Diagnosis type", y = "TFF1 concentration (ng/ml)", fill = "Diagnosis") +
  scale_fill_manual(values = c("firebrick2", "springgreen", "steelblue1"),
                    labels = c("Control","Benign","PDAC"))
```

Urinary levels of protein REG1A that may be associated with pancreas regeneration. Though this data has only been assessed in 306 patients since one goal of the original study was to assess REG1B vs REG1A. But the combination of both bio-markers might be very interesting to use in machine learning. But, since the data is not present for all entries, this data might be removed to increase accuracy of the machine learning algorithm. For now it's necessary to assess if they can give a clear hint of PDAC like LYVE1 and TFF1 seem to be able to do.

```
ggplot(DF_measured_data, aes(x=diagnosis, y=REG1A, fill=diagnosis)) +
  geom_boxplot() +
  scale_x_discrete(breaks = 1:3, labels=c("Control","Benign","PDAC")) +
  labs(x = "Diagnosis type", y = "REG1A concentration (ng/ml)", fill = "Diagnosis") +
  scale_fill_manual(values = c("firebrick2", "springgreen", "steelblue1"),
                    labels = c("Control","Benign","PDAC"))
```

```
## Warning: Removed 284 rows containing non-finite values (stat_boxplot).
```

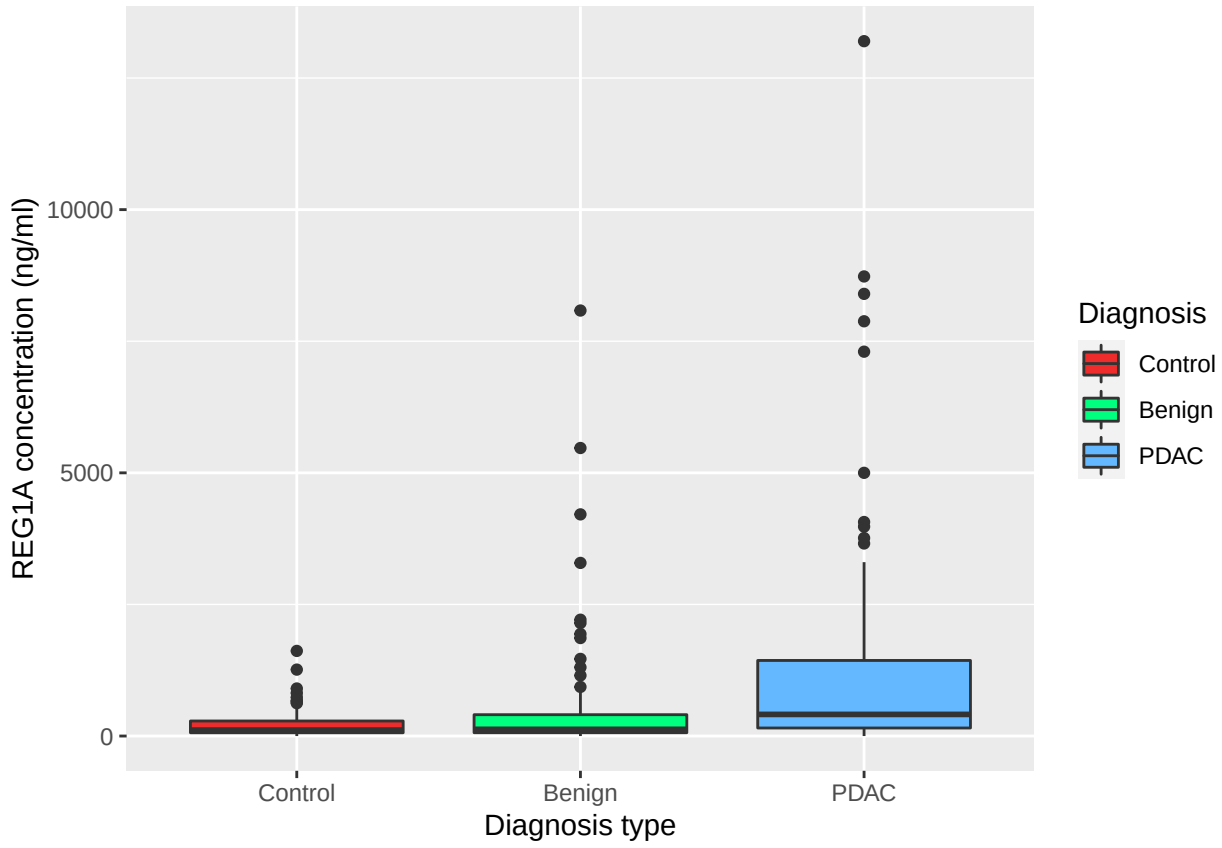

Figure 5: Concentration of REG1A in patients of different diagnosis.

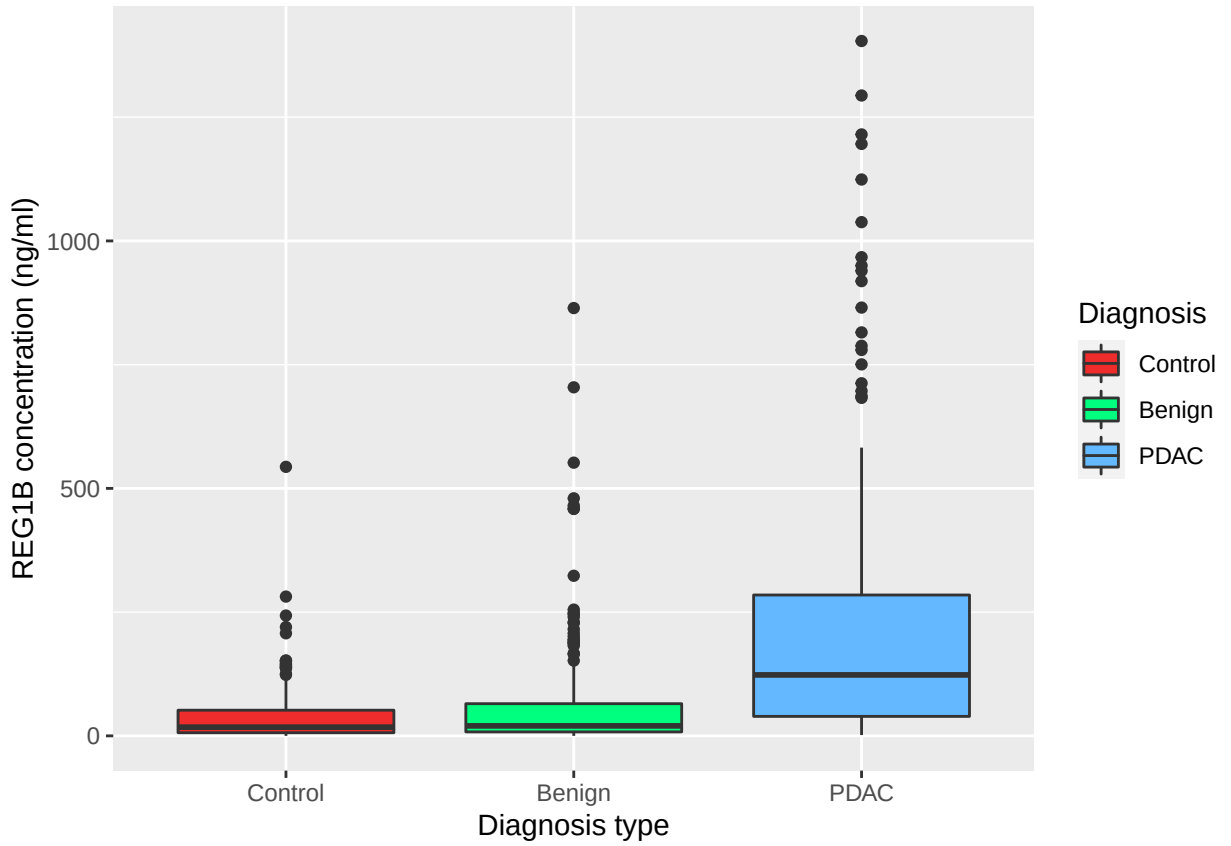

Figure 6: Concentration of REG1B in patients of different diagnosis.

```
ggplot(DF_measured_data, aes(x=diagnosis, y=REG1B, fill=diagnosis)) +
  geom_boxplot() +
  scale_x_discrete(breaks = 1:3, labels=c("Control","Benign","PDAC")) +
  labs(x = "Diagnosis type", y = "REG1B concentration (ng/ml)", fill = "Diagnosis") +
  scale_fill_manual(values = c("firebrick2", "springgreen", "steelblue1"),
                    labels = c("Control","Benign","PDAC"))
```

With REG1B the mean and Q3 of concentrations in PDAC patients are raised higher than with REG1A. There also seem to be more outliers. Although there are more entries in the REG1B group, fig 7. suggests that REG1B is slightly more effective than REG1A.

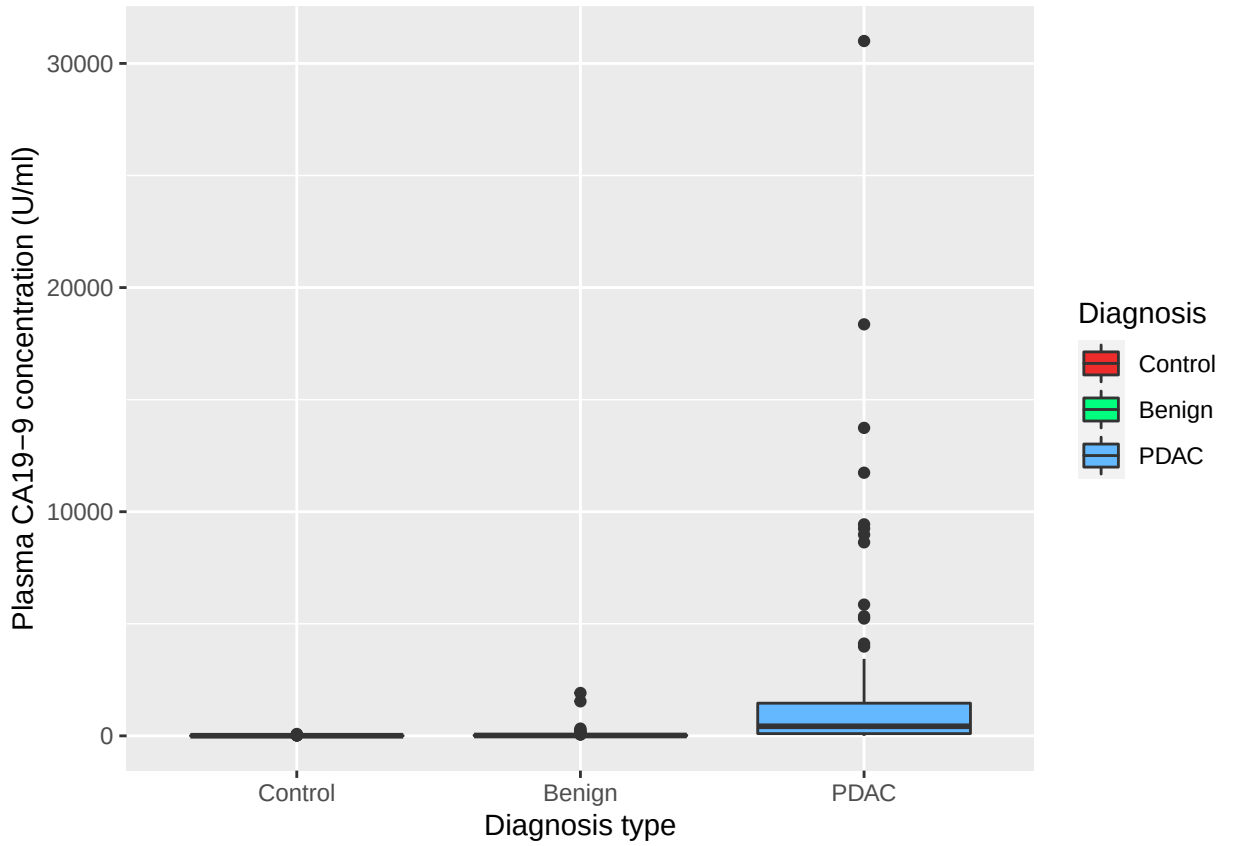

Figure 7: Concentration of plasma CA19-9 in patients of different diagnosis.

The blood plasma levels of CA 19–9 monoclonal antibody is often elevated in patients with pancreatic cancer. Though once again, this has only been assessed in 350 patients (another goal of the study was to compare various CA 19-9 cutpoints from a blood sample to the model developed using urinary samples). Using plasma CA19-9 as the a maker seems therefore like a good idea.

```
ggplot(DF_measured_data, aes(x=diagnosis, y=plasma_CA19_9, fill=diagnosis)) +
  geom_boxplot() +
  scale_x_discrete(breaks = 1:3, labels=c("Control","Benign","PDAC")) +
  labs(x = "Diagnosis type", y = "Plasma CA19-9 concentration (U/ml)", fill = "Diagnosis") +
  scale_fill_manual(values = c("firebrick2", "springgreen", "steelblue1"),
                    labels = c("Control","Benign","PDAC"))
```

```
## Warning: Removed 240 rows containing non-finite values (stat_boxplot).
```

There are 240 rows missing of the plasma CA19-9 entry. This means that this entry is unfavourable to use with a machine learning algorithm. The data will still be used however. As it could prove very beneficial in combination with other markers. The danger is that the machine learning algorithm simply looks if this data is known, if so, the patient probably has PDAC. That much can be concluded from fig 8., in order to keep the algorithm as accurate as possible CA19-9 could be weighed very lightly.

It would seem that the concentrations of the original study's bio-markers all have higher concentrations in PDAC patients, suggesting all of them are possible red flags for PDAC. But what's also important is that just because a patient has a high concentration of I.E. LYVE1, doesn't mean that patient has PDAC. The patient could be afflicted by another form of cancer. For this the usage of CA19-9 can play an important role. High levels of CA19-9 in combination with I.E. LYVE1 could spell PDAC for the patient quite clearly.

Cleaning the data

To enable more accurate results with machine learning, some of the data will have to be modified or removed. Some attributes might not relevant to the research question, while others can vary too much creating a bias in the algorithm.

A number of attributes in the dataset have been recorded for administrative purposes. Things like patient or entry ID and cohort (what subject group the patient was part of) and the sample origin will not be needed for the machine learning algorithm. These data is not relevant to either the individual patient or to the research question. A new dataset wil be created and irrelevant attributes will be removed from this new dataset.

```
cleaned_DF <- subset(DF_measured_data, select = -c(sample_id, patient_cohort, sample_origin))
head(cleaned_DF)
```

```
##   age sex diagnosis stage benign_sample_diagnosis plasma_CA19_9 creatinine
## 1  33   F         1                                        11.7    1.83222
## 2  81   F         1                                          NA    0.97266
## 3  51   M         1                                         7.0    0.78039
## 4  61   M         1                                         8.0    0.70122
## 5  62   M         1                                         9.0    0.21489
## 6  53   M         1                                          NA    0.84825
##        LYVE1     REG1B     TFF1    REG1A
## 1 0.89321920  52.94884 654.2822 1262.000
## 2 2.03758500  94.46703 209.4882  228.407
## 3 0.14558890 102.36600 461.1410       NA
## 4 0.00280488  60.57900 142.9500       NA
## 5 0.00085956  65.54000  41.0880       NA
## 6 0.00339300  62.12600  59.7930       NA
```

The 'age' attribute tells the age of the patient. As discussed in figure 2, this attribute might seem very relevant. With age, the risk of health complications increases. Therefore a bias toword older people might occur. As figure 2 shows, the group that has been diagnosed with PDAC mainly consists of men of middle age and older. However, the control group and the benign diagnosis group both have patients' age reletivly equally distributed. The ML algorithm could therefore create a bias toword older men, rather than looking at the markers.

To confirm, the summarized data shown has been filtered based on the diagnosis. Control group:

```
summary(filter(cleaned_DF, diagnosis==1)$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   48.50   57.00   56.33   63.00   89.00
```

Benign diagnosis:

```
summary(filter(cleaned_DF, diagnosis==2)$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0    44.0    54.0    54.7    66.0    82.0
```

PDAC diagnosis:

```
summary(filter(cleaned_DF, diagnosis==3)$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   59.00   67.00   66.18   74.00   88.00
```

When the data is filtered to only include entries at with a PDAC diagnosis, the youngest patient of the PDAC diagnosis is 29, which is an outlier. The 1st quartile is at 59 years old, with the 3rd quartile at 74. It's also evident that the PDAC diagnosis group consists of older people. With the 1st, 2nd and 3rd as well as the

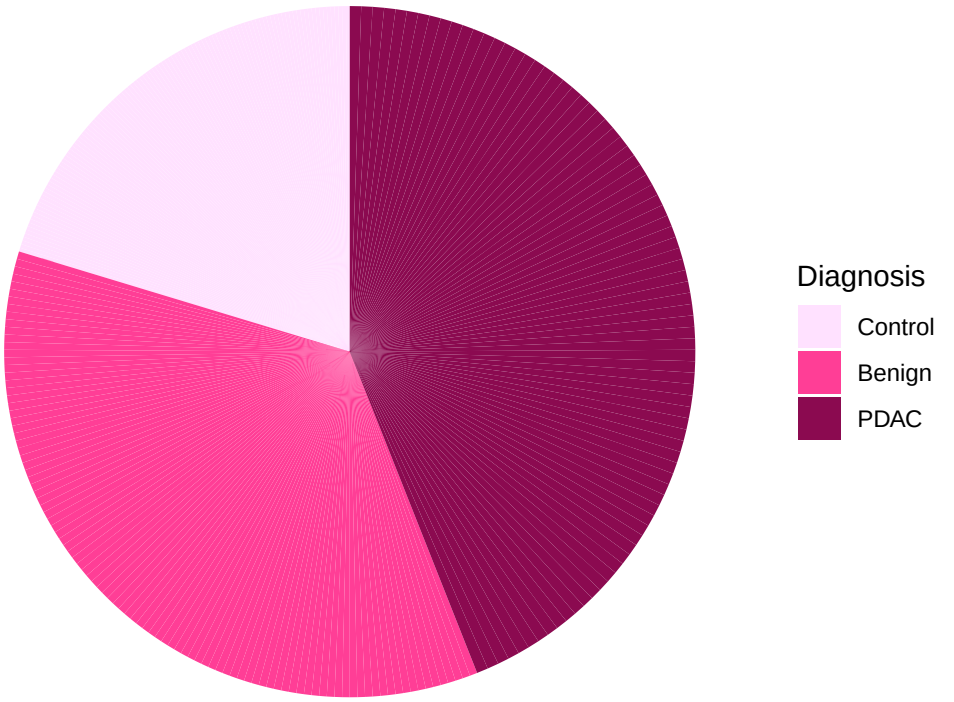

Figure 8: Division of diagnosis of female subjects.

mean being 10+ years older. An argument can be made to drop the age attribute, since there seems to be correlation between the age and sex attributes. However, because the risk of diseases like cancer increases with age, the age attribute will be left in. If need be, this attribute can be removed later if the machine learning algorithm focuses on this variable too much.

The sex attribute notes whether the entry has a male or female body. Because men tend to wait more before seeking medical help, the PDAC diagnosis group contains a lot of male entries, as shown below.

```
ggplot(filter(cleaned_DF, sex=="F"), aes(x="", y=diagnosis, fill=diagnosis)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  scale_fill_manual(values = c("thistle1", "violetred1", "deeppink4"), labels = c("Control", "Benign",
  guides(fill=guide_legend(title="Diagnosis"))
```

```
ggplot(filter(cleaned_DF, sex=="M"), aes(x="", y=diagnosis, fill=diagnosis)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  scale_fill_manual(values = c("cadetblue1", "dodgerblue2", "slateblue4"), labels = c("Control", "Benig
  guides(fill=guide_legend(title="Diagnosis"))
```

The majority of male entries has been diagnosed with PDAC, while this does not hold true for the female entries. Because this correlation likely has roots in psychological and cultural reasons rather than biological reasons (2), this attribute will be dropped.

```
cleaned_DF <- subset(cleaned_DF, select = -sex)
head(cleaned_DF)
```

```
##   age diagnosis stage benign_sample_diagnosis plasma_CA19_9 creatinine
## 1  33         1                                        11.7    1.83222
```

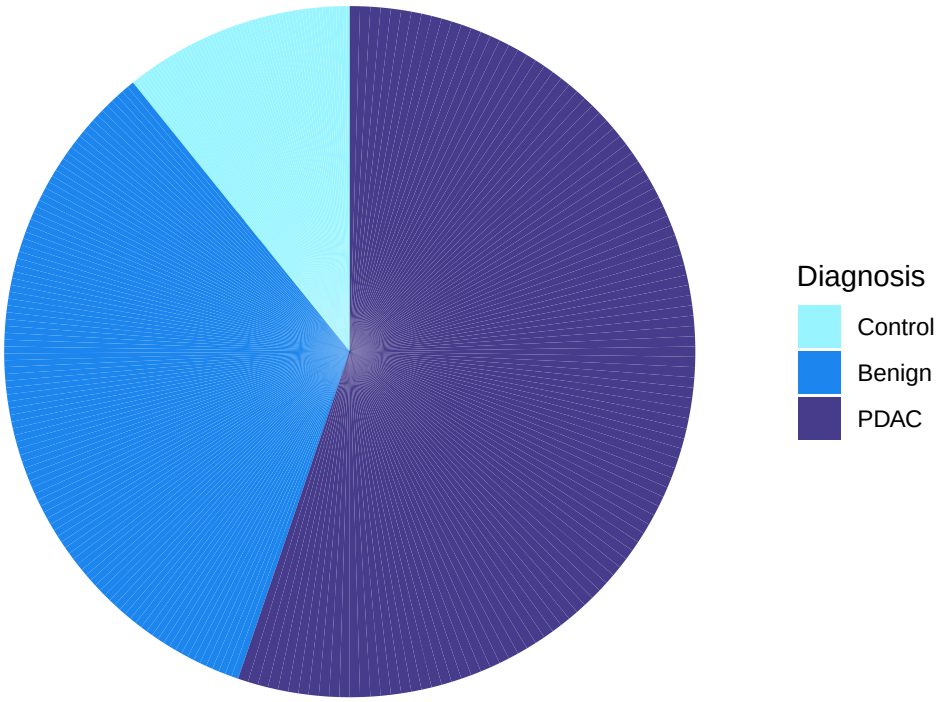

Figure 9: Division of diagnosis of male subjects.

```
## 2  81         1                                          NA    0.97266
## 3  51         1                                         7.0    0.78039
## 4  61         1                                         8.0    0.70122
## 5  62         1                                         9.0    0.21489
## 6  53         1                                          NA    0.84825
##        LYVE1     REG1B     TFF1    REG1A
## 1 0.89321920  52.94884 654.2822 1262.000
## 2 2.03758500  94.46703 209.4882  228.407
## 3 0.14558890 102.36600 461.1410       NA
## 4 0.00280488  60.57900 142.9500       NA
## 5 0.00085956  65.54000  41.0880       NA
## 6 0.00339300  62.12600  59.7930       NA
```

Next is diagnosis. The data for this attribute is complete without missing, or NA values. It is also necessary to differentiate between the results. The machine learning algorithm can look at what concentrations are common for a certain diagnosis, and base the results of this.

Stage is an interesting attribute that tells how far along a patient with PDAC is. Because this data is only recorded for patients who have been diagnosed with PDAC, the relevancy to the research question might seem vague. The stage attribute can potentially say something about the concentrations of the markers, which may be beneficial. The counter-argument is that the goal of this research is to find the most telling markers, not to find out whether these markers are able to diagnose the stage the cancer is in. Besides that, the goal of the algorithm is to determine a diagnosis of the patient based of the concentrations of each marker, and in doing so find out which marker is the most effective. The stage of PDAC patients is not relevant to that goal in this particular time. Also, the patient must already have been diagnosed with PDAC in order to get this information in the first place. It is not something one could feed to an machine learning algorithm to diagnose the patient. Therefore the attribute has been dropped.

```
cleaned_DF <- subset(cleaned_DF, select = -stage)
head(cleaned_DF)
```

```
##   age diagnosis benign_sample_diagnosis plasma_CA19_9 creatinine
## 1  33         1                                 11.7    1.83222
## 2  81         1                                   NA    0.97266
## 3  51         1                                  7.0    0.78039
## 4  61         1                                  8.0    0.70122
## 5  62         1                                  9.0    0.21489
## 6  53         1                                   NA    0.84825
##        LYVE1     REG1B     TFF1    REG1A
## 1 0.89321920  52.94884 654.2822 1262.000
## 2 2.03758500  94.46703 209.4882  228.407
## 3 0.14558890 102.36600 461.1410       NA
## 4 0.00280488  60.57900 142.9500       NA
## 5 0.00085956  65.54000  41.0880       NA
## 6 0.00339300  62.12600  59.7930       NA
```

The Benign sample diagnosis marks for those with a benign, non-cancerous diagnosis, what the diagnosis is. It can be argued that this data can be useful in that it might help train the algorithm to recognize certain concentrations of the bio-markers for another condition, like gallstones or chronic pancreatitis. However, the data does not have enough entries to train the algorithm with enough accuracy. For now, learning what markers, or combination of markers clearly point to either a benign or PDAC diagnosis will do. The benign sample diagnosis attribute has been dropped for that reason.

```
cleaned_DF <- subset(cleaned_DF, select = -benign_sample_diagnosis)
head(cleaned_DF)
```

```
##   age diagnosis plasma_CA19_9 creatinine      LYVE1     REG1B     TFF1
## 1  33         1          11.7    1.83222 0.89321920  52.94884 654.2822
## 2  81         1            NA    0.97266 2.03758500  94.46703 209.4882
## 3  51         1           7.0    0.78039 0.14558890 102.36600 461.1410
## 4  61         1           8.0    0.70122 0.00280488  60.57900 142.9500
## 5  62         1           9.0    0.21489 0.00085956  65.54000  41.0880
## 6  53         1            NA    0.84825 0.00339300  62.12600  59.7930
##      REG1A
## 1 1262.000
## 2  228.407
## 3       NA
## 4       NA
## 5       NA
## 6       NA
```

REG1B and TFF1 show a number of entries that can range across orders of magnitude. To reign in the large differences, the data in REG1B and TFF1 will be log(2) transformed.

```
cleaned_DF[, 6:7] <- log(cleaned_DF[6:7], 2)
```

The attributes that note the concentration of plasma CA19-9, creatinine, LYVE1, REG1B, TFF1 and REG1A will be kept because they are crucial for answering the research question. This does not mean the entries are complete however. Beause the minimum detectable level of CA19-9 from the original samples was 0.3 U/ml using the FLUOstar Omega Microplate Reader, there are some values which are unknown and have been marked with 'NA' instead. However, we can safely deduce that the concentrations of CA19-9 must be within 0 and 0.3. High concentrations of CA19-9 are often found in cancer patients, and should therefore be regarded as a red flag. There are however missing values with confirmed PDAC entries as well. Luckily, the cencentrations of creatinine, LYVE1, REG1B and TFF1 follow a trend of having higher concentrations in PDAC patients compared to the benign and control groups, it is possible to use the average distance to the standard deviation of the know marker concentrations of each entry to deduce what the level of CA19-9 or REG1A should or could be.

```r
#Dit werkt nog niet 100%, maar bijna wel. De logic klopt, er gaat alleen ergens iets fout met het vervan

fill_na <- function(entry){

  #Because the entry DF is numeric in nature, the entries will be converted to numeric as well.
  entry <- as.numeric(entry)

  crea_mean <- mean(filter(cleaned_DF, diagnosis==entry[2])$creatinine)
  dist_crea <- (crea_mean / 100) * as.numeric(entry[4])

  LYVE1_mean <- mean(filter(cleaned_DF, diagnosis==entry[2])$LYVE1)
  dist_LYVE1 <- (LYVE1_mean / 100) * as.numeric(entry[5])

  REG1B_mean <- mean(filter(cleaned_DF, diagnosis==entry[2])$REG1B)
  dist_REG1B <- (REG1B_mean / 100) * as.numeric(entry[6])

  TFF1_mean <- mean(filter(cleaned_DF, diagnosis==entry[2])$TFF1)
  dist_TFF1 <- (TFF1_mean / 100) * as.numeric(entry[7])

  avr_dist <- mean(c(dist_crea, dist_LYVE1, dist_REG1B, dist_TFF1))
  mean_CA19_9 <- mean(filter(cleaned_DF, diagnosis==entry[2])$plasma_CA19_9, rm.na=TRUE)
  mean_REG1A <- mean(filter(cleaned_DF, diagnosis==entry[2])$REG1A, rm.na=TRUE)

  #Calculate the values that may fill in the NA spaces.
  to_fill_CA19_9 <- (mean_CA19_9 / 100) * avr_dist
  to_fill_REG1A <- (mean_REG1A / 100) * avr_dist

  toReturn <- entry

  #Check if plasma_CA19_9 is NA, if so, fill it in.
  toReturn[3] <- ifelse(is.na(toReturn[3]), to_fill_CA19_9, toReturn[3])

  #Check if REG1A is NA, if so, fill it in.
  toReturn[8] <- ifelse(is.na(toReturn[8]), to_fill_REG1A, toReturn[8])

  return(toReturn)
}

#test <- apply(cleaned_DF, MARGIN = 2, fill_na)
```

Results

The data has been cleaned of any administrative entries, like the patient cohort, sample id and sample origin. As these are entries that are important to note, but are not very useful to feed the machine learning algorithm. The values have been checked for any attributes whose entries can differ by order of magnitude. These attributes were found to be the concentrations of REG1B and TFF1 respectivly. In order to avoid any future complications arrising from this difference, these entries have been log2 transformed, reigning in the large difference and making the values easier to work with. Futher inspection of the data revealed that the attribues logging the concentrations of plasma CA19-9 and REG1A have several NA entries. This is likely due to the minimum concentration neccesary for optical density equipment to detect the concentration. Also, because this data is collected from humans, by humans, it's not beyond the realm of possibility that mistakes were made during the collection of data. Either way, NA entries are in the data. In order to remove these entries,

a function has been written that calculates the average distance of the known and complete concentrations (creatinine, LYVE1, REG1B, TFF1) to the mean in a percentage. Because these concentrations are higher in PDAC group than in the benign group, and higher in benign group than the control group, the function takes the diagnosis into account. The function determines the mean of the average difference between the mean of the concentration and the noted data, to determine how far the entries concentrations are from the norm. Then it assigns a value to the NA entry of plasma_CA19_9 and REG1A based on this calculation. In a shorter explination, if an entry's concentrations on average are around 70% of the total average concentration of any particular marker in the entry's diagnosis group, the NA value of IE CA19-9 will also be 70% of the known CA19-9 concentration.

Discussion and conclusion

The data has been pretty well organized. Though proper care in weighing possible NA values is very important as these can have a distorting effect on the ML algorithm. The concentrations of the markers seem higher in the benign and PDAC groups. The data does contain decent amount of NA entries, which makes finding an accurate to the reseach without cleaning the NA values, a difficult if not impossible task. Some correlations between attributes are very clear, even without directly comparing them. Having used these correlations to fill in the NA values is a risky manouvre however. The idea of generating data, and using said generated data to teach an algorithm how to recognize diseases, can potentially go very wrong. By generating this data based on what's already known about the entry, and the correlation between the markers, this risk has been minimalized as much as possible. Simply using the mean of each attribute and diagnosis group to fill in the NA values would make that attribute less trustworthy. With the fill_na function, it's possible to fill the NA values of each attribute without risking contaminating an entry with low concentrations with a high concentration of IE CA19-9. The other option would of course be to remove any entries with NA values. But this would severly cut down the data that is to be be fed to the machine learning algorithm. Less data means less accurate results. So the debate is really about wheter it's more accurate to use less data, or to generate the unknow data. By generating values to replace the NA entries by using known values and limits, hopefully the data is as accurate as can be. Which in turn means the machine learning algorithm is as accurate as it can be.

Sources

(1): 'A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study' by S. Debernardi et al. of the Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London. The data used in this EDA has been taken from this source. (https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003489&rev=2#sec008)

(2): 'No man's land: men, illness, and the NHS' by I. Banks of the European Men's Health Forum. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1121551/).