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?

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

```r
#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")
```

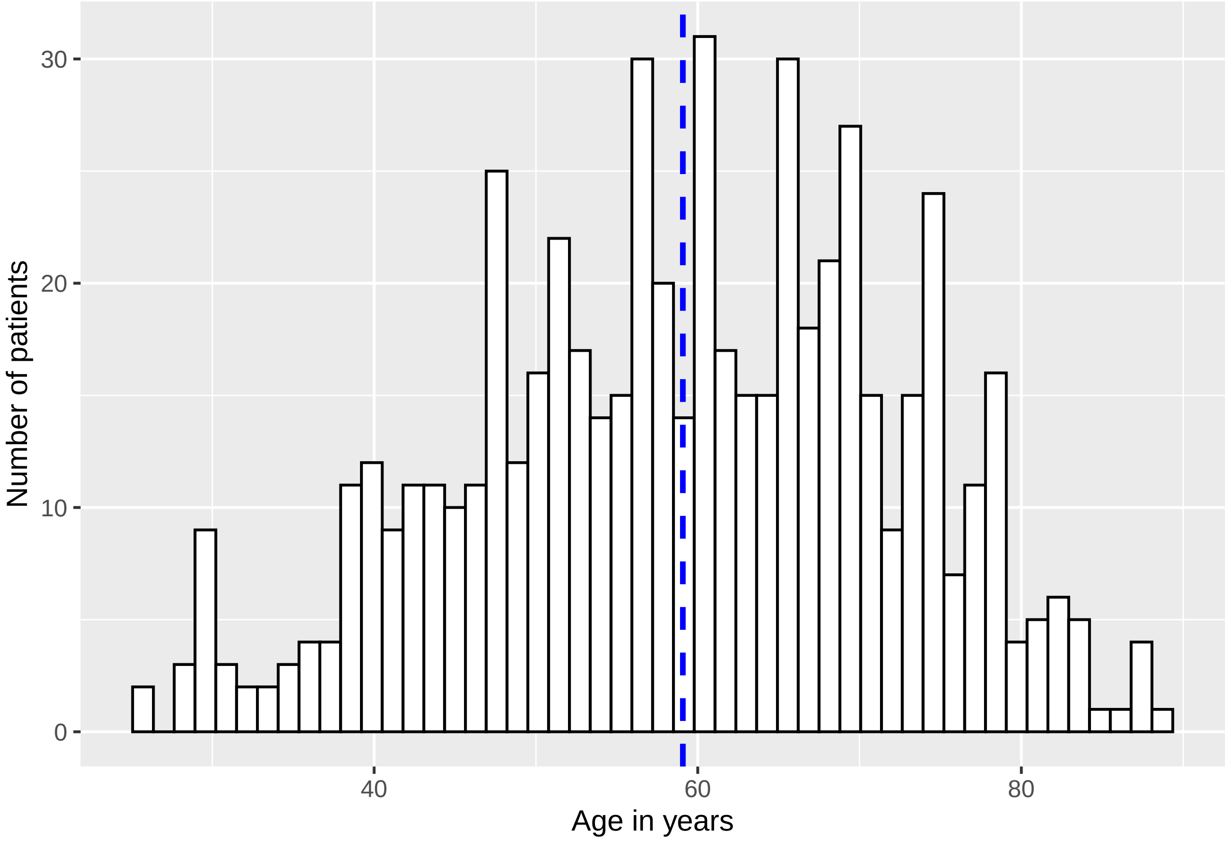

Figure 1 suggest that tested subjects are fairly evenly divided based on age. Though the distribution leans in favour of the elderly. Since the data contains both control, benign and PDAC groups, the above figure can be misleading. As it doesn't account for any group bias.

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

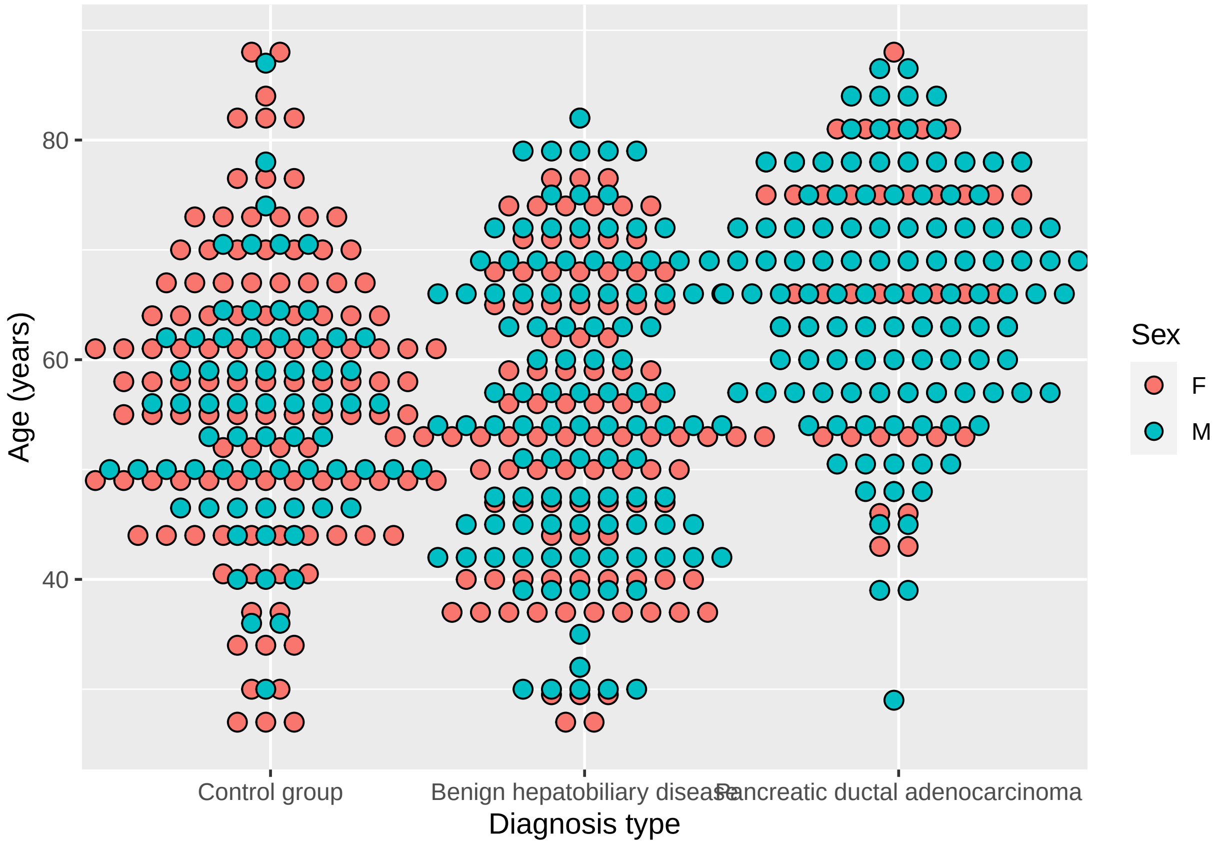

Figure 1: Patients grouped by diagnosis, distributed by age.

```
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.

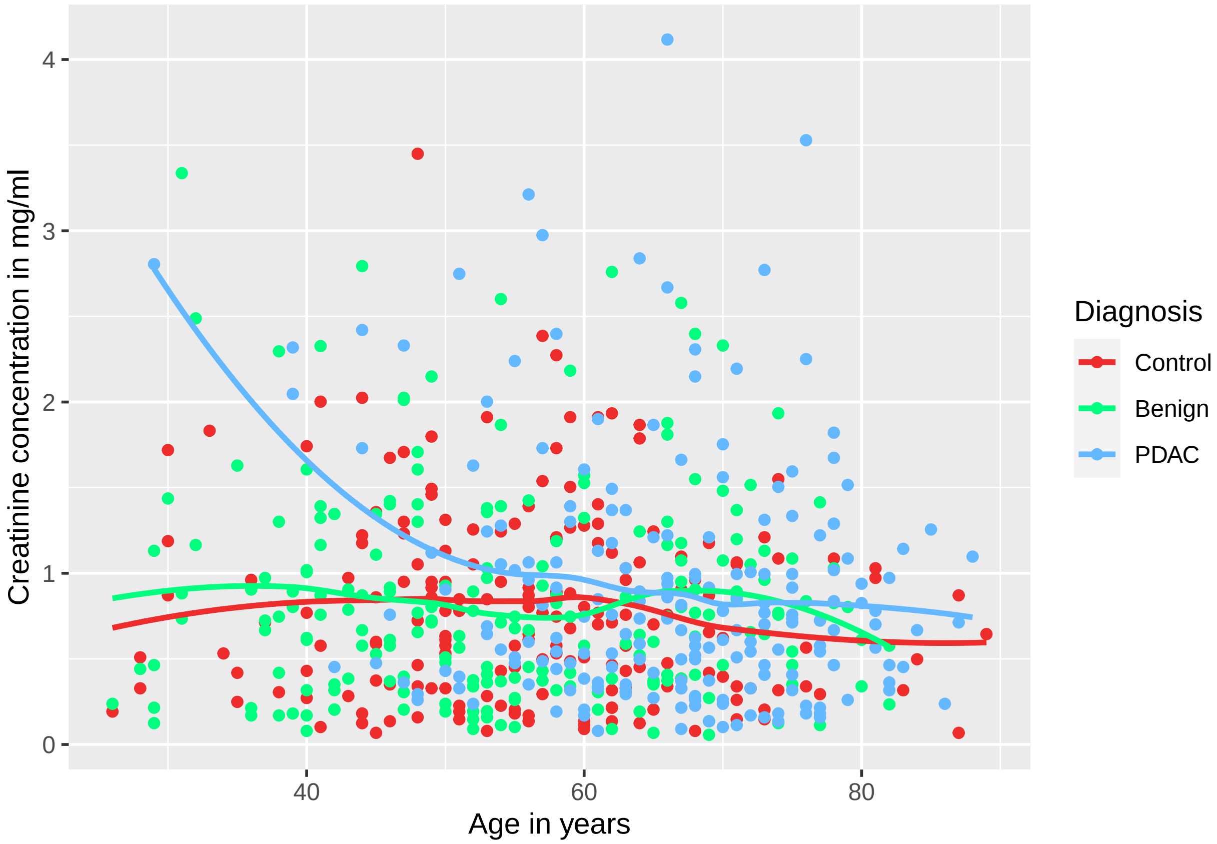

Figure 2: Correlation between age and concentration of creatinine.

```
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.

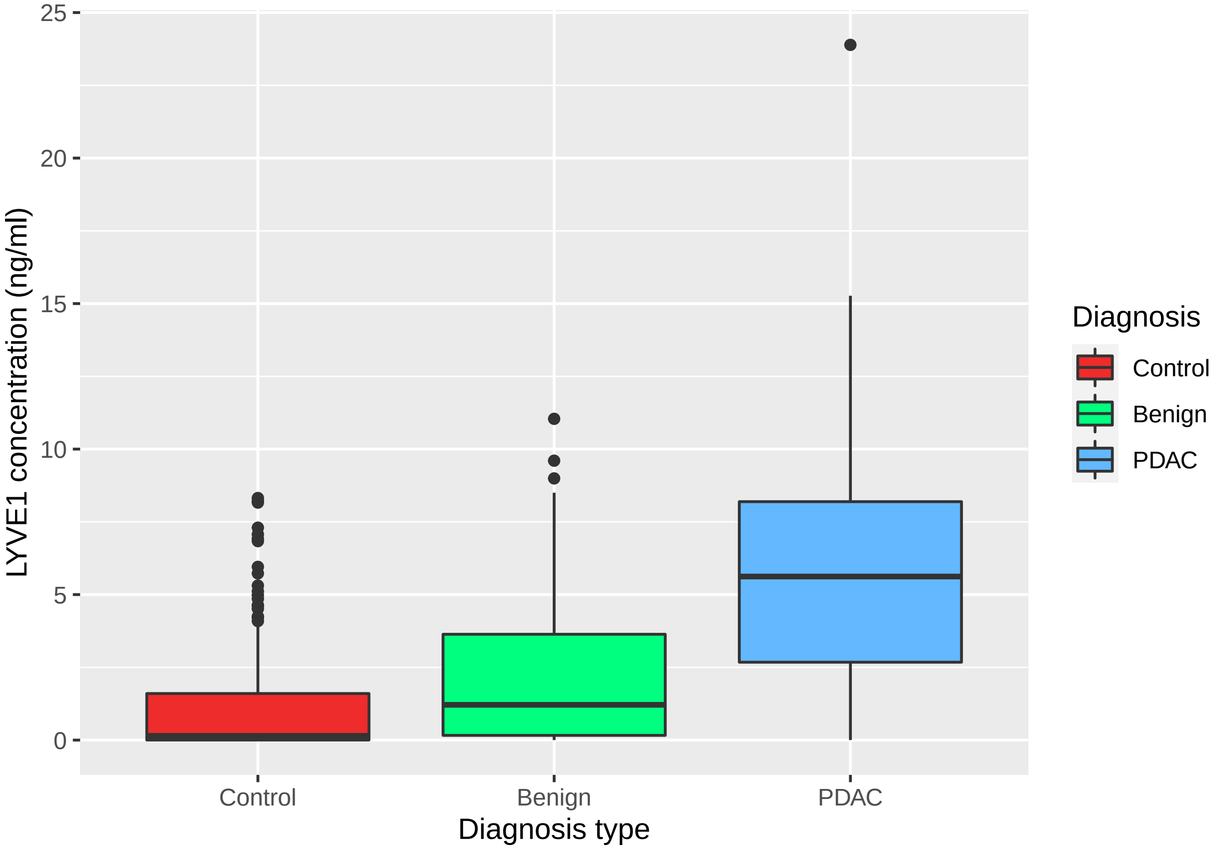

Figure 3: Concentration of LYVE1 in patients of different diagnosis.

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.

```r
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.

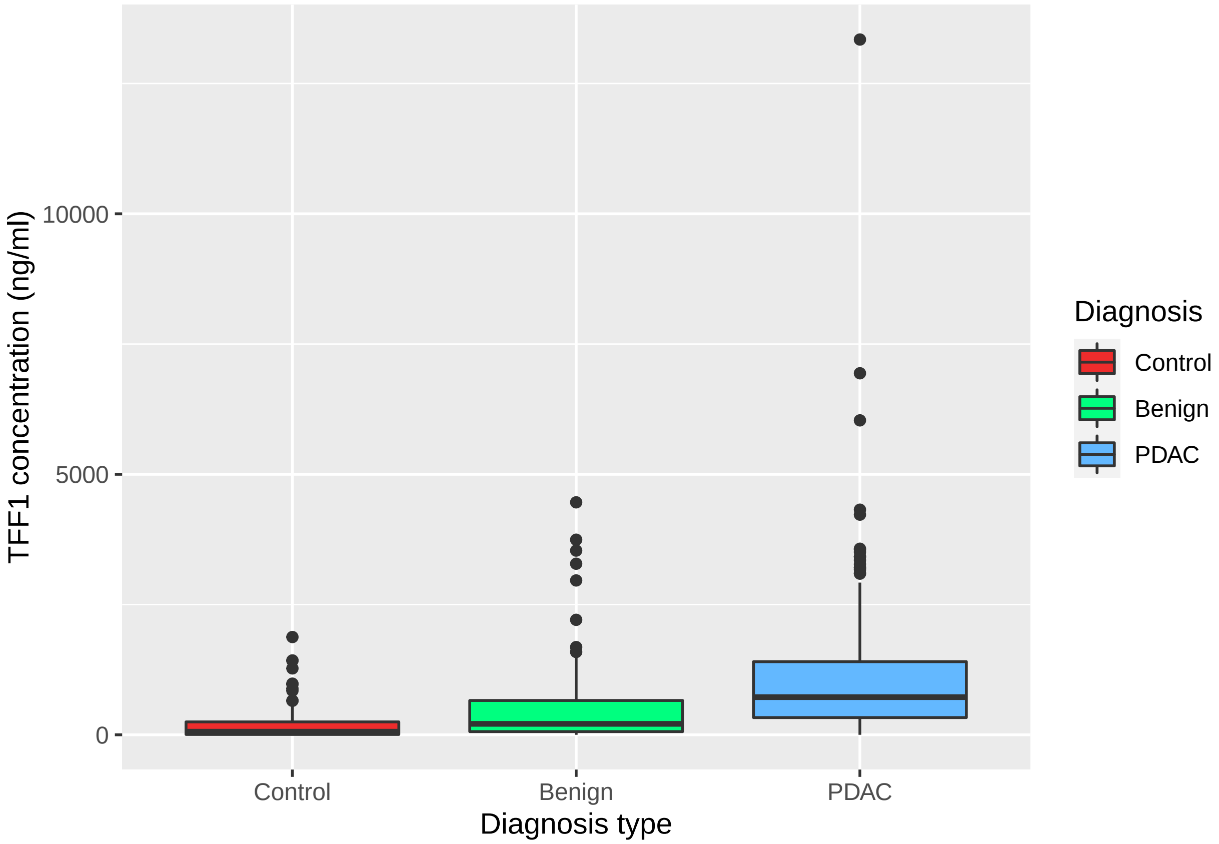

Figure 4: Concentration of TFF1 in patients of different diagnosis.

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).
```

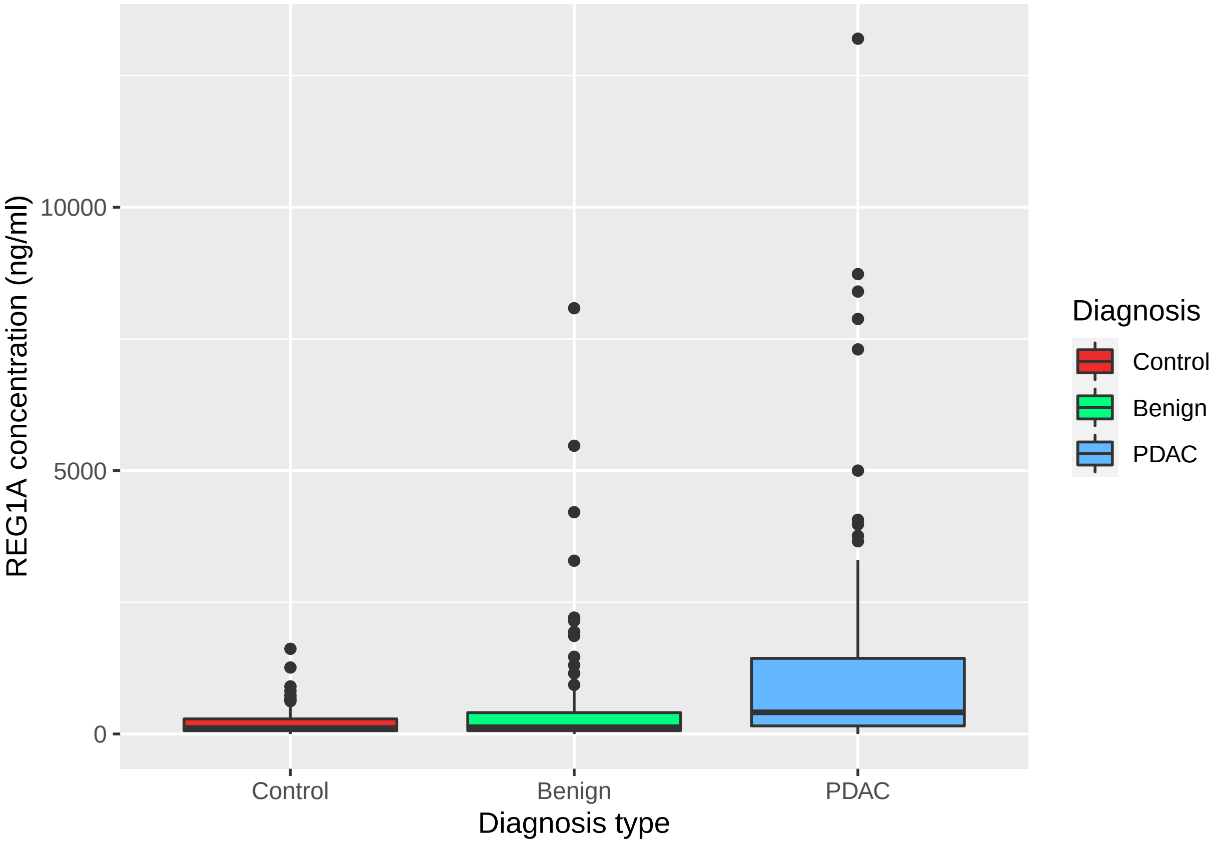

Figure 5: Concentration of REG1A in patients of different diagnosis.

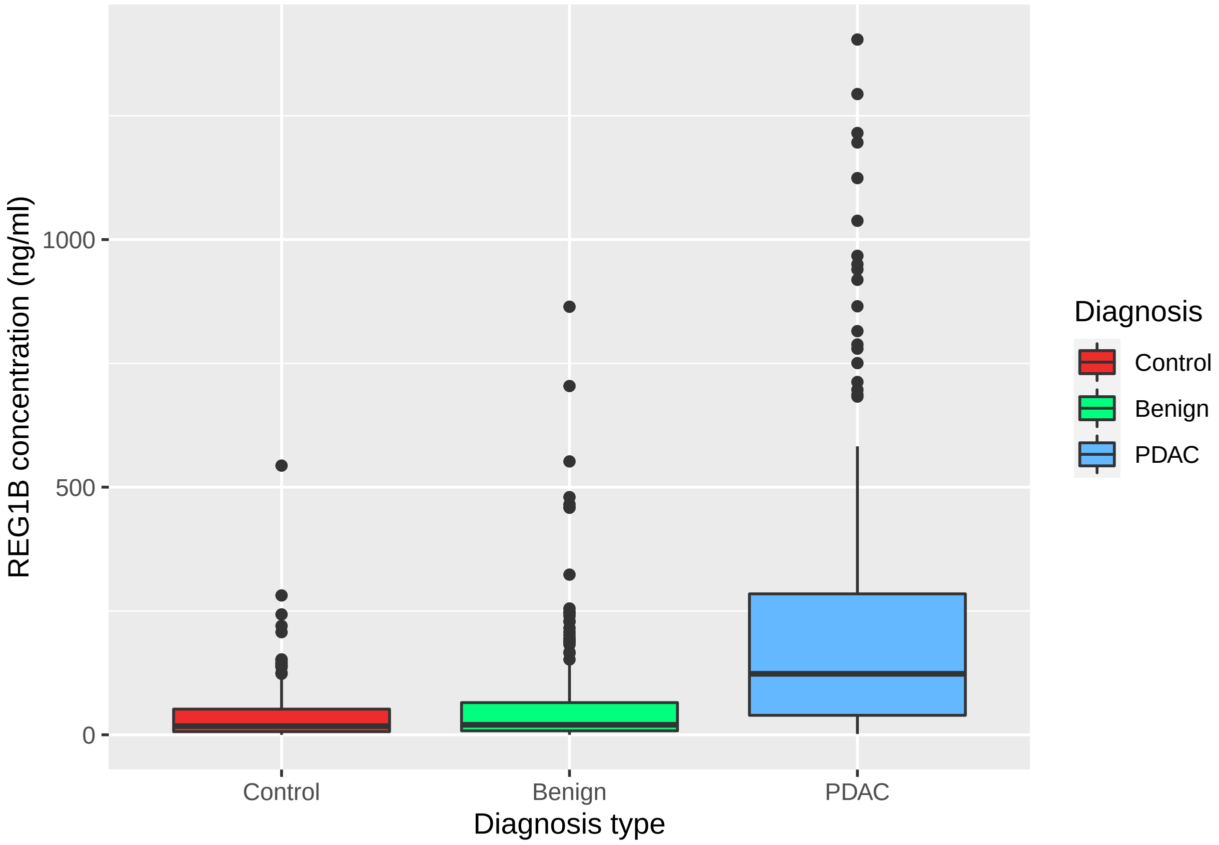

Figure 6: Concentration of REG1B in patients of different diagnosis.

```
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.

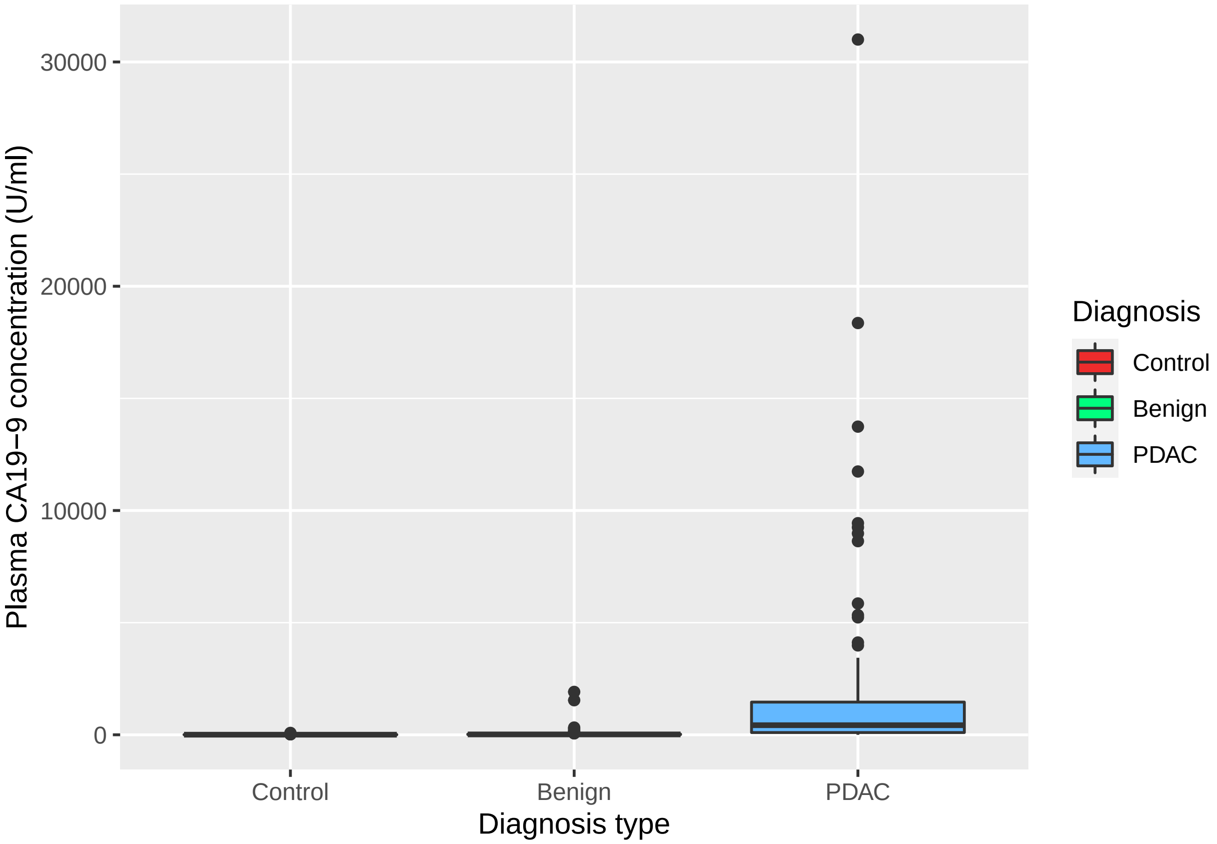

Figure 7: Concentration of plasma CA19-9 in patients of different diagnosis.

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 know, 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 plasma CA19-9 should 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.

Conclusion

The data seems 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 data does not contain large amounts of NA entries, which makes it largely usable. Some correlations between attributes are very clear, even without directly comparing them, but since all bio-markers measured are ones already involved with PDAC and other forms of cancer, they must be weighed with caution to prevent bias in the final algorithm.

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/).