Exploratory Data Analysis - PanC and T2D

Edwina Rossi

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important step when working with large scale data sets, such as the epidemiological data set used in this project. The aim of the EDA is to investigate the data and summarise its main characteristics through the means of plots and summary statistics. The main results from the EDA can be found in the final report “Results” section and in “Appendix D - Exploratory Data Analysis”

Counting number of Cases and Controls

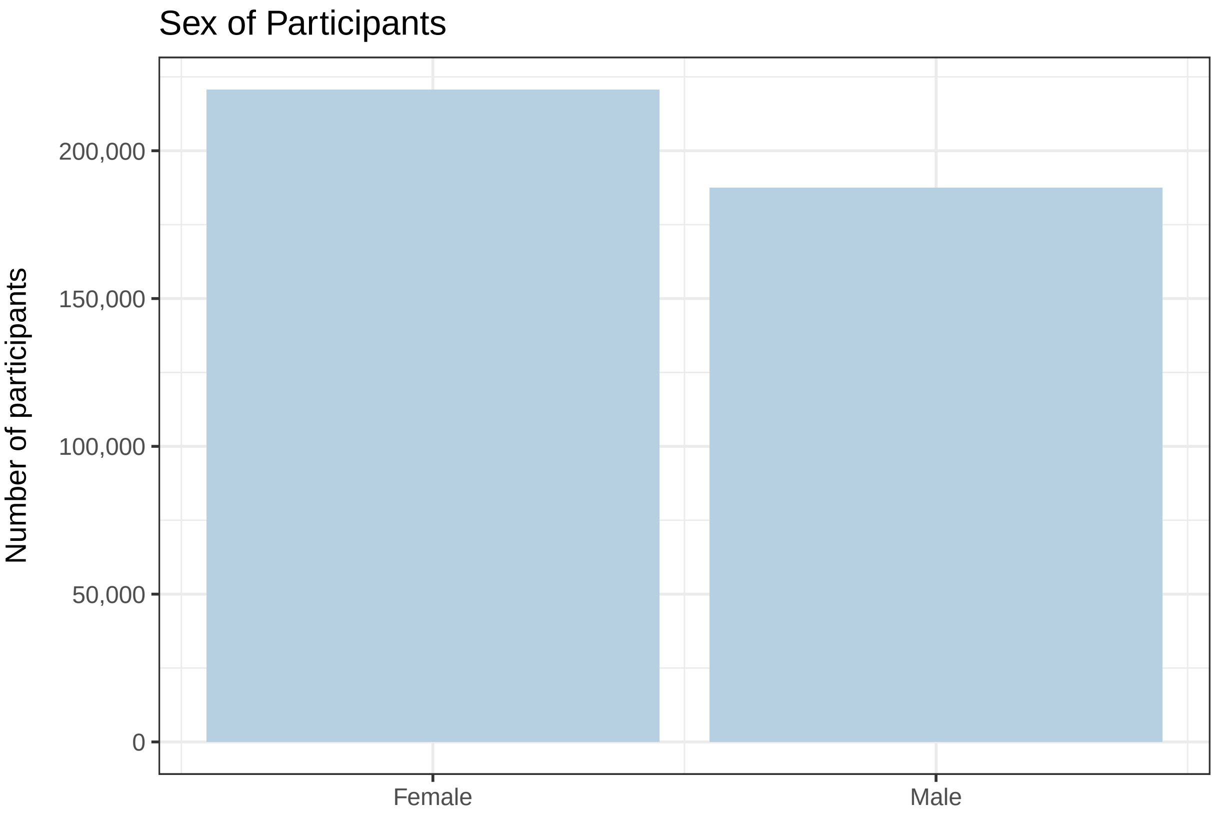
```
# Note that T2D = E11, and PanC = C25

## T2D and PANC
#table of PanC and T2D cases and controls
with(epi_data, addmargins(table(T2D, C25)))
```

```
##      C25
## T2D        0      1    Sum
##   0   374072   1147 375219
##   1    32597    384  32981
##   Sum 406669   1531 408200
```

Study Participant Information Plots

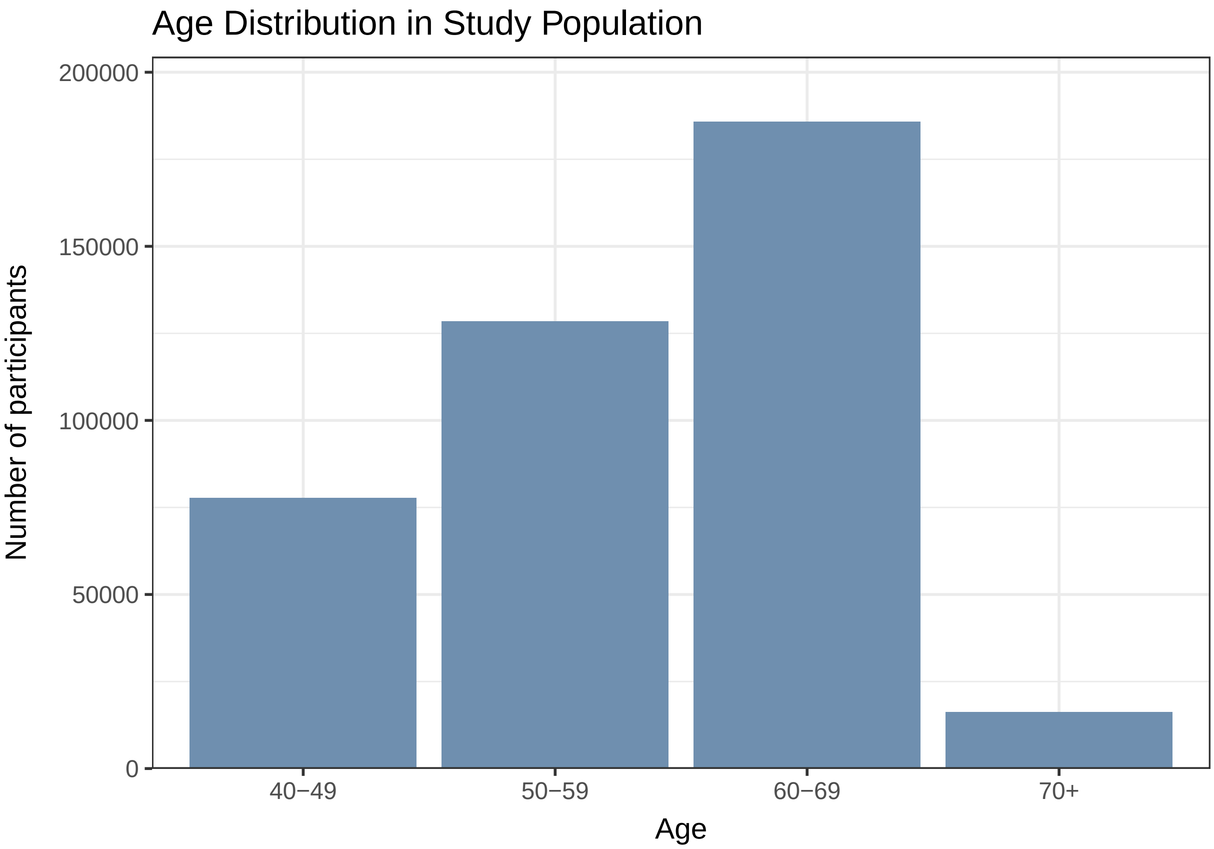
Sex

The distribution of sex (as determined by genotyping) in the epidemiological data is as follows:



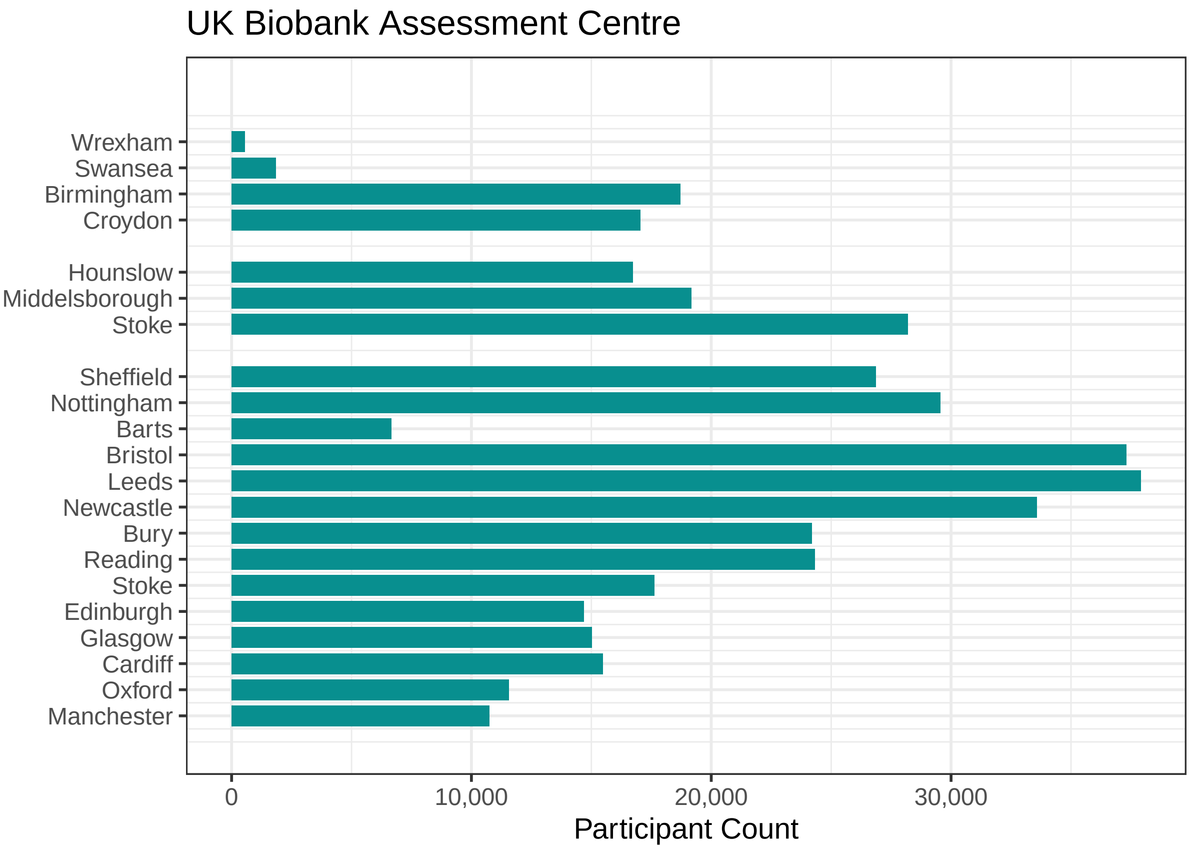
Age Distribution

The participants of the UK Biobank study were all aged 40-69 at time of recruitment. The following bar chart illustrates the distribution of participants across age groups 40-49, 50-59 and 60-69 at time of recruitment, and also stratified by sex.



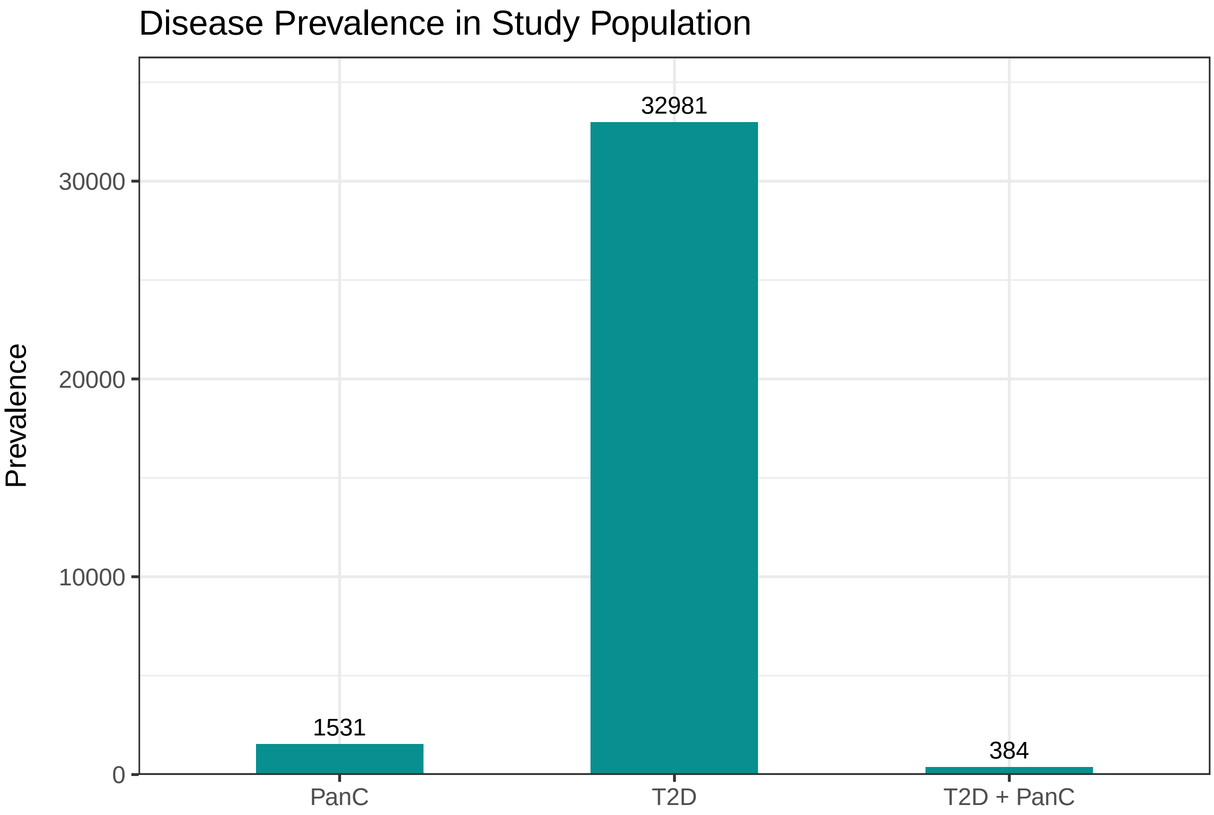
UK Biobank Assessment Centre

The participants in the UK Biobank were assessed at 22 different locations across the UK. The following histogram shows the distribution of participants across the 22 assessment centres:



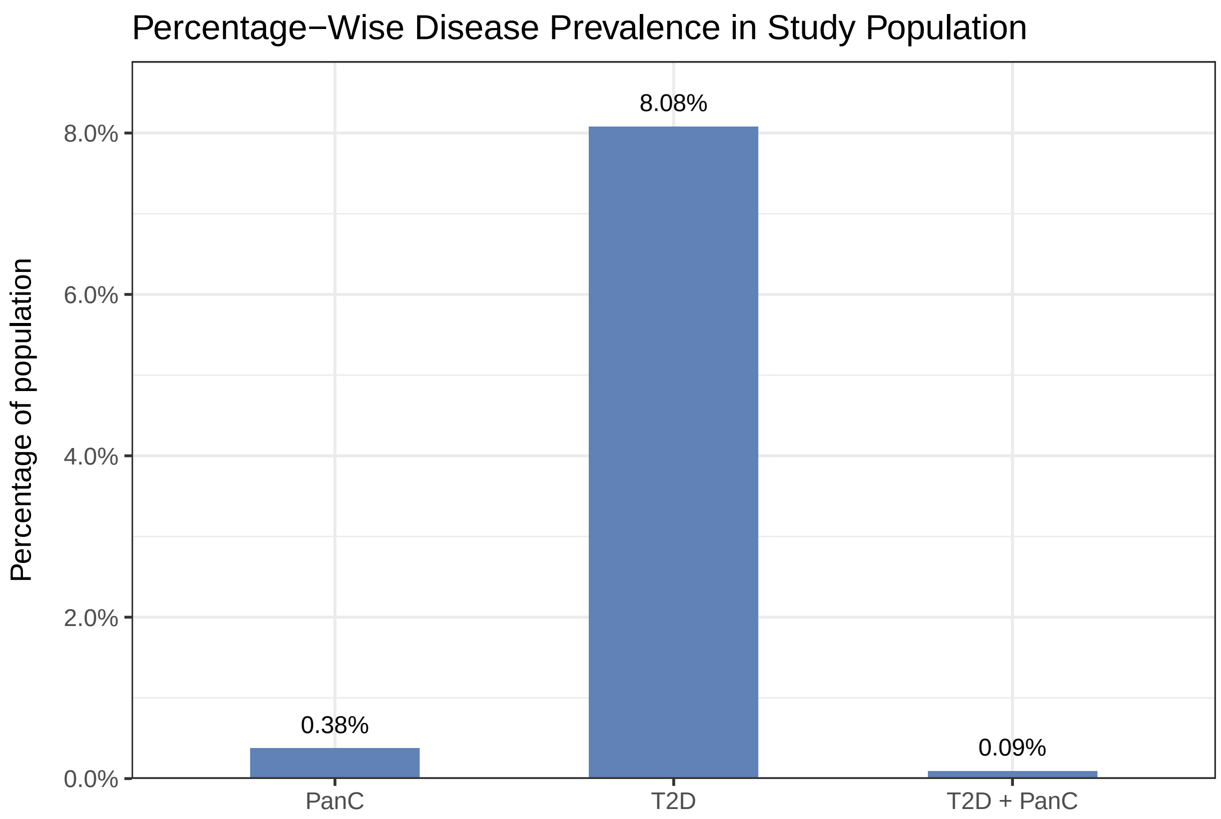
Disease Prevalence

The below plot shows the prevalence of T2D, PanC and T2D + PanC diagnoses within the UK BB study population, as well as the number of cases of T2D + PanC where T2D was diagnosed > 1 after PanC.



Percentage-Wise Disease Prevalence

The below plots show disease prevalence of PanC, T2D and T2D + PanC in percentage terms.



Are people with PANC more likely to have a T2D diagnosis, or the other way around?

To understand whether T2D patients are more likely to develop PanC, or whether PanC patients are more likely to develop T2D, we may compare the prevalence of the two diseases in different groups.

```
#table of PanC and T2D cases and controls
with(epi_data, addmargins(table(T2D, C25)))
```

```
##      C25
## T2D        0      1    Sum
##   0   374072   1147 375219
##   1    32597    384  32981
##   Sum 406669   1531 408200
```

```
#table of PanC and T2D cases and controls
with(epi_data, addmargins(table(C25, T2D)))
```

```
##      T2D
## C25        0      1    Sum
##   0   374072  32597 406669
##   1     1147    384   1531
##   Sum 375219  32981 408200
```

```
# Percentage table of PanC and T2D cases and controls
with(epi_data, prop.table(table(T2D, C25), margin = 1)*100)
```

```
##    C25
## T2D          0          1
##   0 99.6943119  0.3056881
##   1 98.8356933  1.1643067
```

```
with(epi_data, prop.table(table(C25, T2D), margin = 1)*100)
```

```
##    T2D
## C25        0        1
##   0 91.98439  8.01561
##   1 74.91835 25.08165
```

```
#Chi-squared test
with(epi_data, chisq.test(T2D, C25))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  T2D and C25
## X-squared = 595.85, df = 1, p-value < 2.2e-16
```

We observe that participants with a T2D diagnosis are ~3 times as likely to also have a PanC diagnosis, compared to participants with out T2D. Equally, we also observe that participants with a PanC diagnosis are ~3 times as likely to also have a T2D diagnosis, compared to participants without PanC

The Chi-squared test between T2D and PANC gives a highly significant p-value, indicating that there indeed is a statistically significant relationship between T2D and PANC.

Comparison to Disease Prevalence in UK Population

To understand whether the disease prevalence in the study population is representative of the disease prevalence in the actual UK population, we may compare the findings herein to data on PanC and T2D prevalence in the UK population.

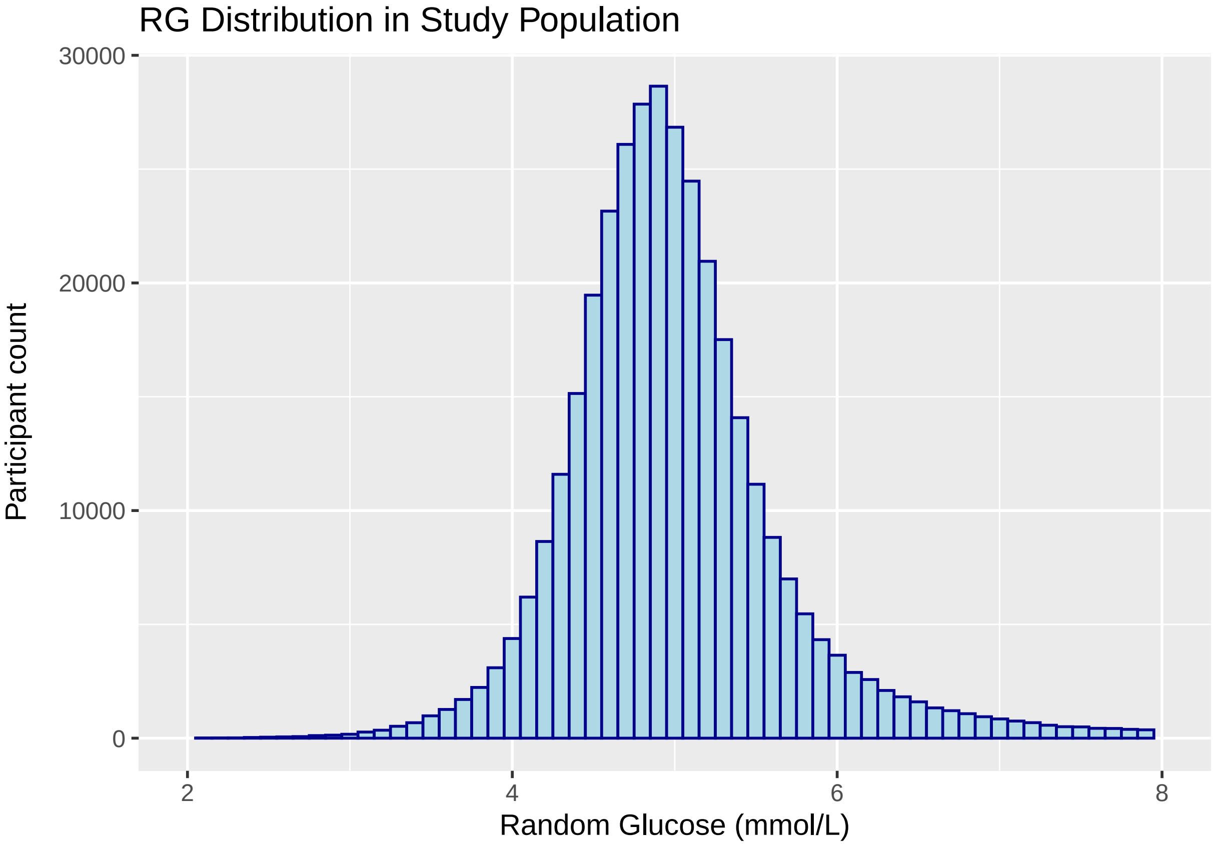
According to data from Cancer Research UK, 1 in 53 UK males and 1 in 57 UK females will be diagnosed with pancreatic cancer during their lifetime. On average, this means that roughly 1 in every 55 person is likely to develop pancreatic cancer. If we take the UK population to be 67.22 million (per 2020), this equates to about 1.82%. This figure is almost 5 times higher than in this project's study population. However, it should be noted that 47% of all new pancreatic cancer cases occur in people aged 85 to 89 (Cancer Research UK), and that this project's study population has a lower average age (at recruitment).

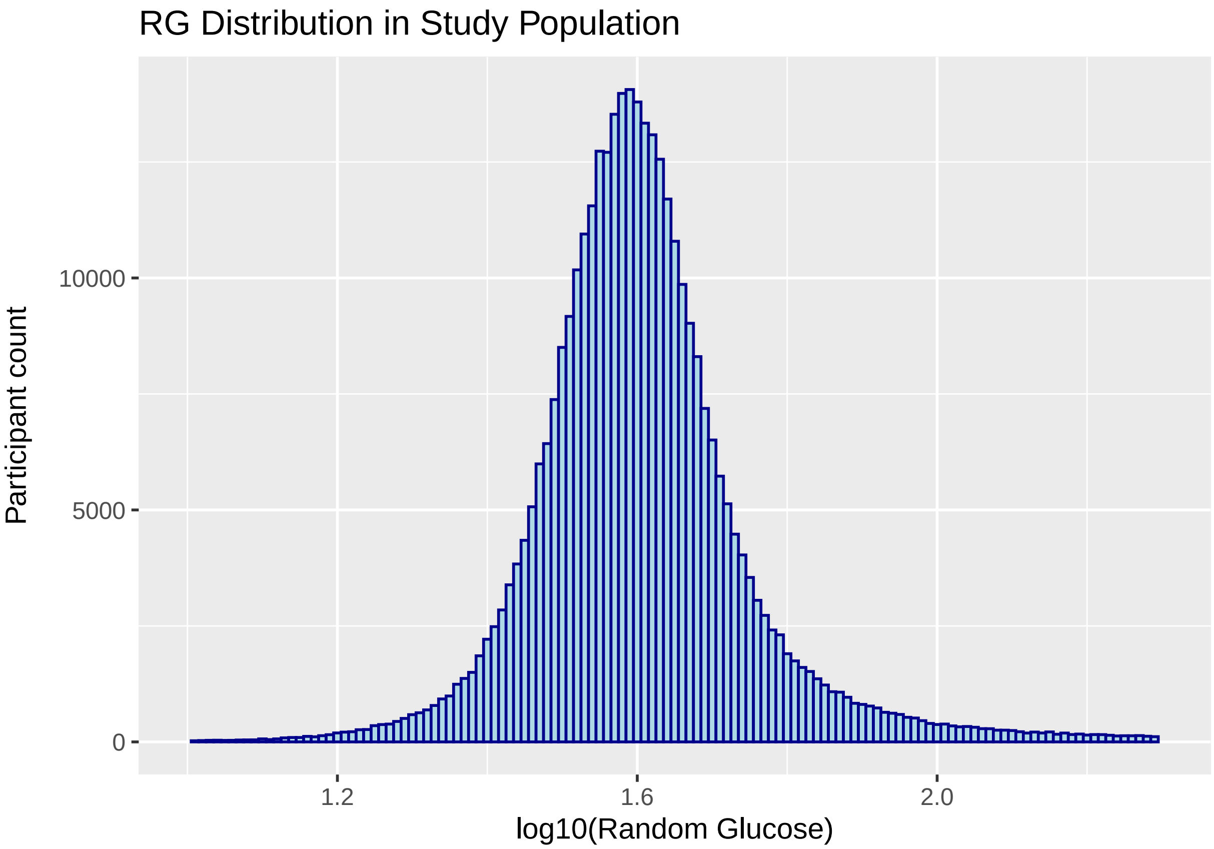
With regards to the prevalence of T2D, data from diabetes.org indicate that 1 in 10 over 40s in the UK currently live with diabetes. Knowing that all subjects in the study population are also over 40 years of age, we can compare disease prevalence in this project's study population to the prevalence in the actual UK population in the over 40s. We note that just over 8% of this project's study population have a diabetes diagnosis, compared to the 10% of people living with diabetes in the UK. Though these figures are relatively similar, there might be cases of undiagnosed diabetes in this project's population, explaining why the prevalence of T2D is a bit lower than in the average population.

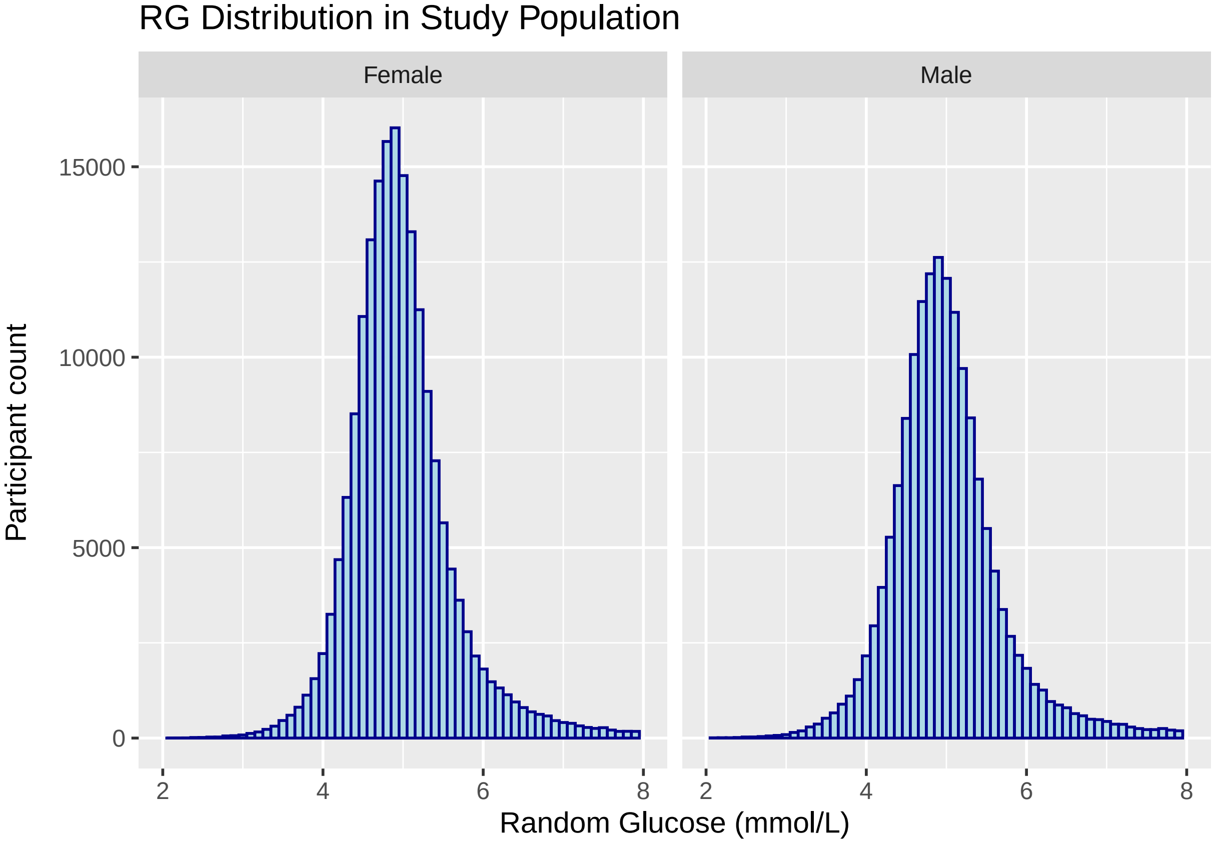
Random Glucose

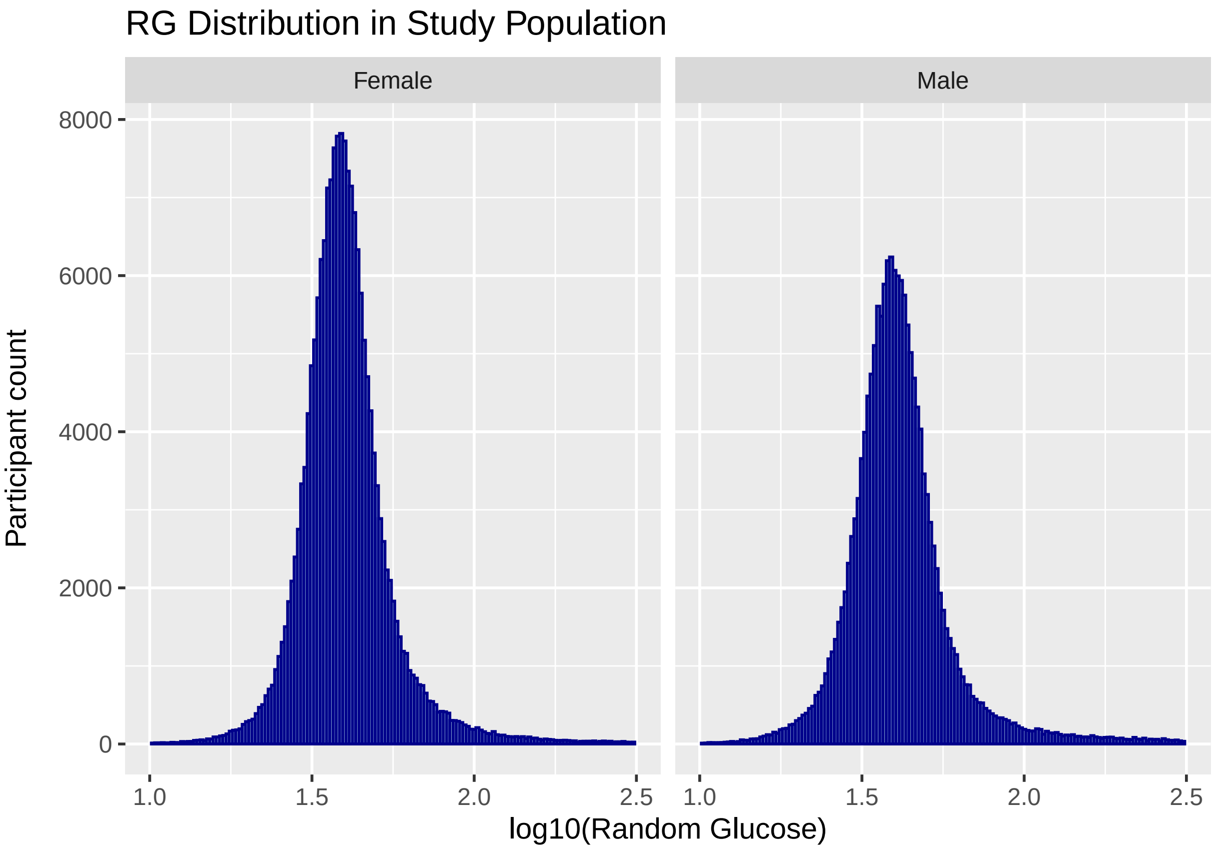
Random Glucose (RG) is a marker of blood glucose values, taken at a random time-point without the patient needing to fast before the blood sample is drawn. It is a common diagnostic marker of diabetes, with values

> 11.1 mmol/L being indicative of T2D. The below plots shows the distribution of RG levels in the study population.




RG Distribution in Study Population
Participant count
10000
5000
0
1.2
1.6
2.0
log10(Random Glucose)

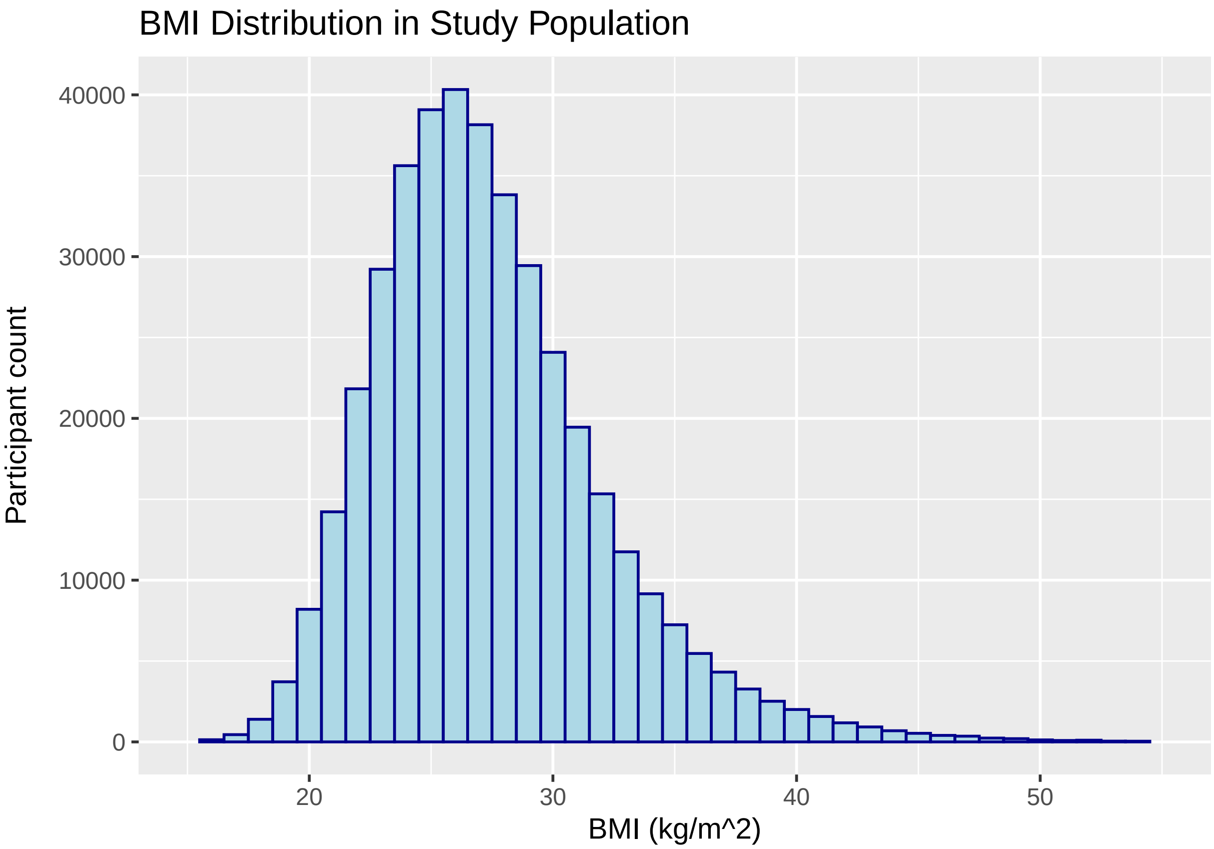

RG Distribution in Study Population
Female
Male
Participant count
15000
10000
5000
0
2
4
6
8
2
4
6
8
Random Glucose (mmol/L)

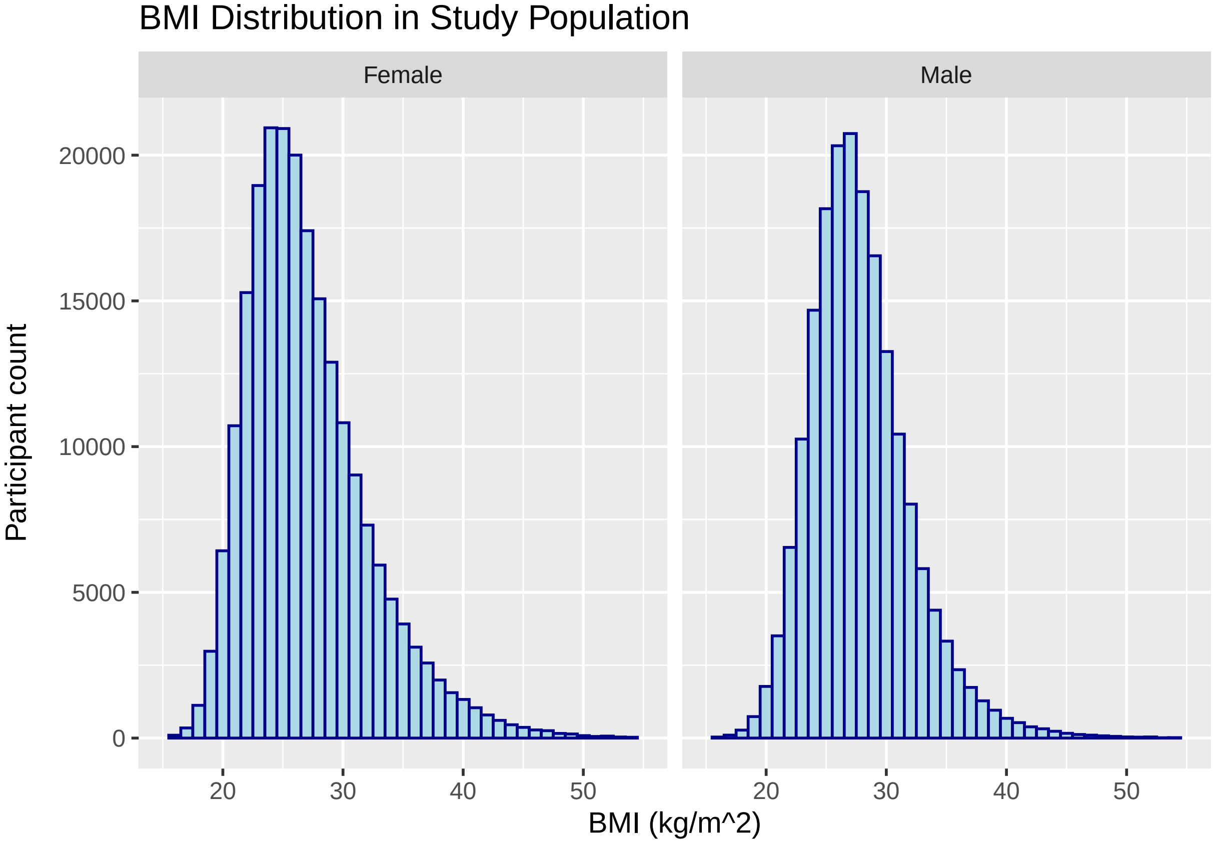


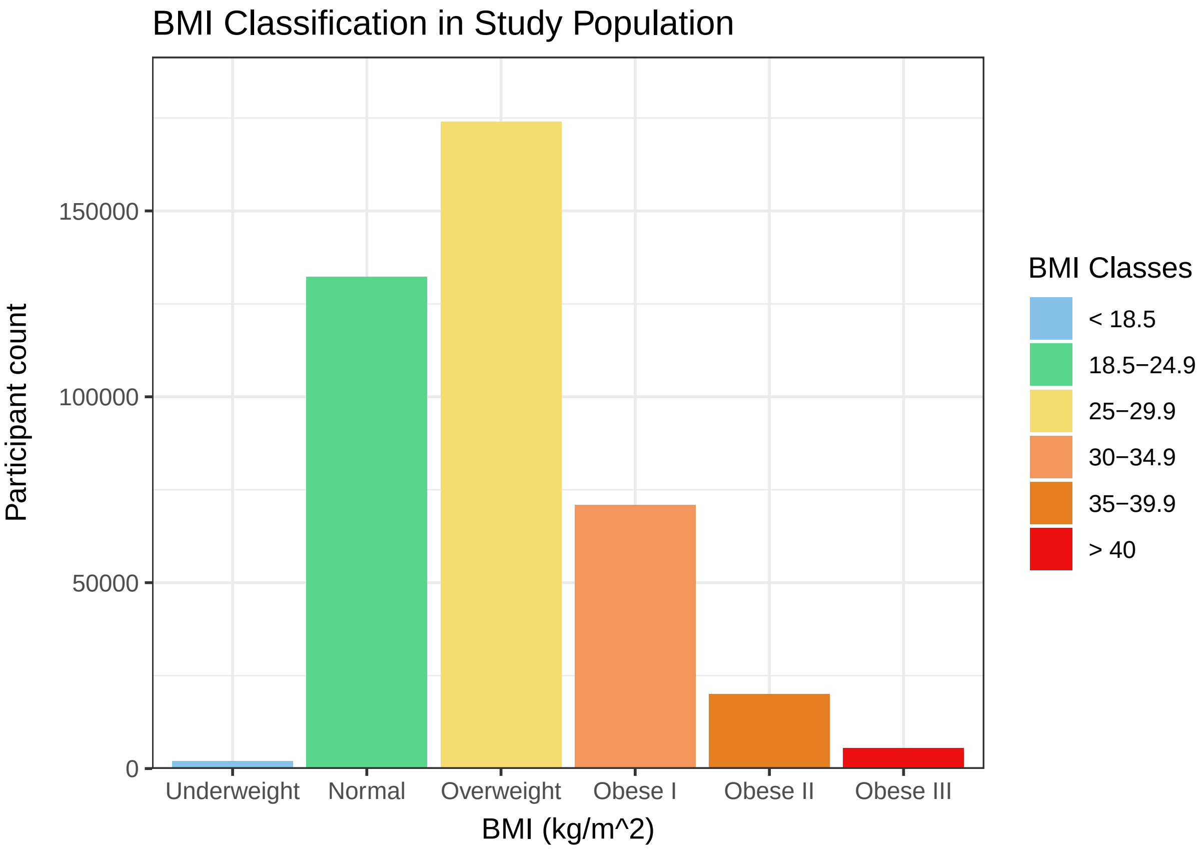
The range of RG levels in the study population is approximately normally distributed.

Body Mass Index (BMI)

Body Mass Index is a measure used to work out whether an individual's weight is healthy. A BMI calculation is performed by dividing an individual's weight (in kg) by their height (in metres) squared (thus, units = kg/mˆ2). BMI values are often stratified into different levels, depending on the what "state of health" the value indicates.


BMI Distribution in Study Population
Participant count
40000
30000
20000
10000
0
20
30
40
50
BMI (kg/m^2)


BMI Distribution in Study Population
Female
Male
Participant count
20000
15000
10000
5000
0
20
30
40
50
20
30
40
50
BMI (kg/m^2)



We observe that the majority of individuals in the study population are in the "overweight" BMI category, with a BMI of 25-29.9 kg/mˆ2.

Summary Statistics

The below calculations of summary statistics are performed in order to provide an overview of key characteristics of the study population. They are calculated for the whole cohort, only for T2D patients, only for PanC patients, and for individuals with both a T2D and PanC diagnosis.

Summary Statistics of Whole Cohort

```
# Mean and SD of RG measurements for whole cohort
mean_RG_all = mean(epi_data$Random.glucose, na.rm = T)
SD_RG_all = sd(epi_data$Random.glucose, na.rm = T)

# Mean and SD of BMI for whole cohort
mean_BMI_all = mean(epi_data$BMI, na.rm = T)
SD_BMI_all = sd(epi_data$BMI, na.rm = T)

# Mean and SD of age for whole cohort
mean_age_all = mean(epi_data$AgeBaseline, na.rm = T)
SD_age_all = sd(epi_data$AgeBaseline, na.rm = T)

# Sex distribution for whole cohort
```

```
female_count_all = sum(epi_data$Sex == 0)
male_count_all = sum(epi_data$Sex == 1)

perc_women = sum(epi_data$Sex == 0, na.rm = T)/nrow(epi_data)
perc_men = sum(epi_data$Sex == 1, na.rm = T)/nrow(epi_data)
```

Summary Statistics for PanC cases

```
# Mean and SD of RG for all participants with a PANC diagnosis
mean_RG_level_PANC = mean(subset(epi_data, epi_data$C25 == 1)$Random.glucose, na.rm = T)
SD_RG_level_PANC = sd(subset(epi_data, epi_data$C25 == 1)$Random.glucose, na.rm = T)

# Mean and SD of BMI for all participants with a PANC diagnosis
mean_BMI_PANC = mean(subset(epi_data, epi_data$C25 == 1)$BMI, na.rm = T)
SD_BMI_PANC = sd(subset(epi_data, epi_data$C25 == 1)$BMI, na.rm = T)

# Mean and SD of age for all participants with a PANC diagnosis
mean_age_PANC = mean(subset(epi_data, epi_data$C25 == 1)$AgeBaseline, na.rm = T)
SD_age_PANC = sd(subset(epi_data, epi_data$C25 == 1)$AgeBaseline, na.rm = T)

# Sex distribution for all participants with a PANC diagnosis
female_count_PANC = sum(subset(epi_data, epi_data$C25 ==1)$Sex == 0)
male_count_PANC = sum(subset(epi_data, epi_data$C25 ==1)$Sex == 1)

perc_women_PANC = sum(subset(epi_data, epi_data$C25 ==1)$Sex == 0)/nrow(subset(epi_data,
                  epi_data$C25 == 1))
perc_men_PANC = sum(subset(epi_data, epi_data$C25 ==1)$Sex == 1)/nrow(subset(epi_data,
                 epi_data$C25 == 1))
```

Summary Statistics for T2D cases

```
# Mean and SD of RG for all participants with a T2D diagnosis
mean_RG_level_T2D = mean(subset(epi_data, epi_data$T2D == 1)$Random.glucose, na.rm = T)
SD_RG_level_T2D = sd(subset(epi_data, epi_data$T2D == 1)$Random.glucose, na.rm = T)

# Mean and SD of BMI for all participants with a T2D diagnosis
mean_BMI_T2D = mean(subset(epi_data, epi_data$T2D == 1)$BMI, na.rm = T)
SD_BMI_T2D = sd(subset(epi_data, epi_data$T2D == 1)$BMI, na.rm = T)

# Mean and SD of age for all participants with a T2D diagnosis
mean_age_T2D = mean(subset(epi_data, epi_data$T2D == 1)$AgeBaseline, na.rm = T)
SD_age_T2D = sd(subset(epi_data, epi_data$T2D == 1)$AgeBaseline, na.rm = T)

# Sex distribution for all participants with a T2D diagnosis
female_count_T2D = sum(subset(epi_data, epi_data$T2D ==1)$Sex == 0)
male_count_T2D = sum(subset(epi_data, epi_data$T2D ==1)$Sex == 1)

perc_women_T2D = sum(subset(epi_data, epi_data$T2D ==1)$Sex == 0)/nrow(subset(epi_data,
                  epi_data$T2D == 1))
```

```
perc_men_T2D = sum(subset(epi_data, epi_data$T2D ==1)$Sex == 1)/nrow(subset(epi_data,
                 epi_data$T2D == 1))
```

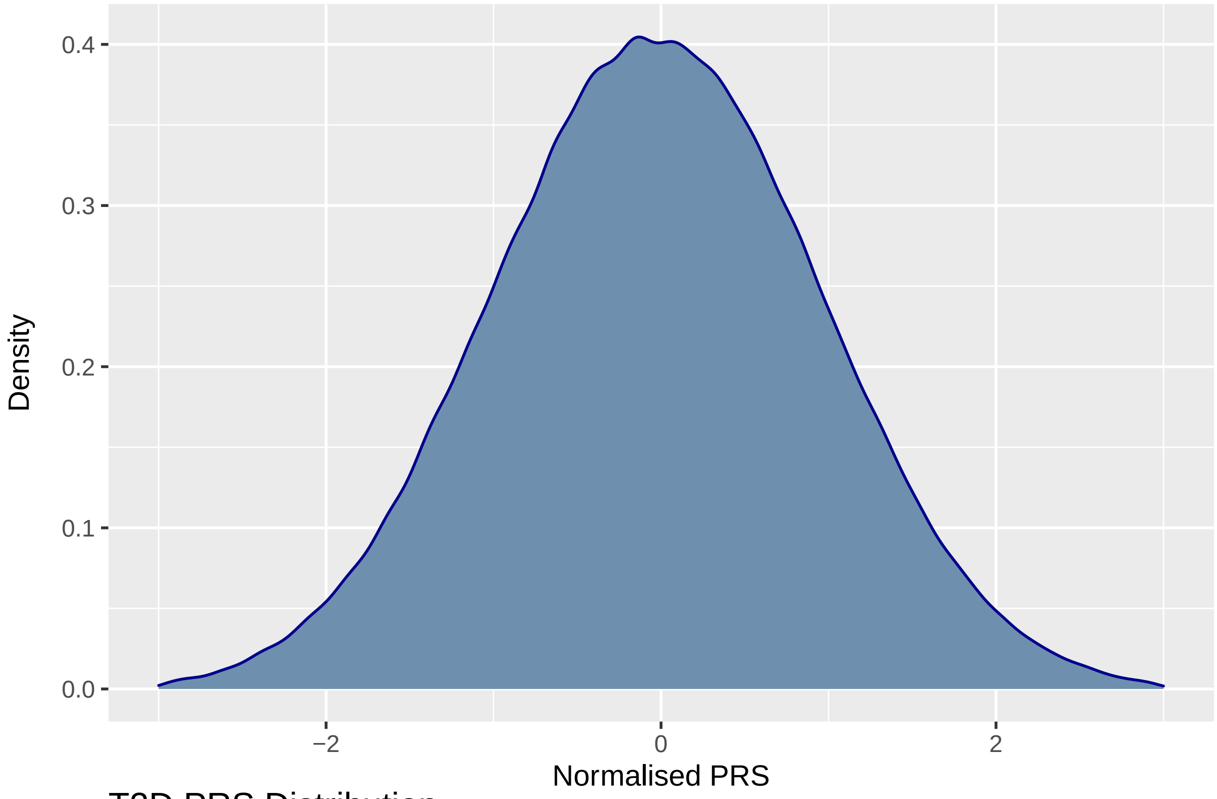
Summary Statistics for T2D + PanC cases

```
mean_RG_level_T2D_PANC = mean(subset(epi_data, epi_data$T2D.PANC == 1)$Random.glucose, na.rm = T)
SD_RG_level_T2D_PANC = sd(subset(epi_data, epi_data$T2D.PANC == 1)$Random.glucose, na.rm = T)

# Mean and SD of BMI for all participants with a PANC diagnosis
mean_BMI_T2D_PANC = mean(subset(epi_data, epi_data$T2D.PANC == 1)$BMI, na.rm = T)
SD_BMI_T2D_PANC = sd(subset(epi_data, epi_data$T2D.PANC == 1)$BMI, na.rm = T)

# Mean and SD of age for all participants with a PANC diagnosis
mean_age_T2D_PANC = mean(subset(epi_data, epi_data$T2D.PANC == 1)$AgeBaseline, na.rm = T)
SD_age_T2D_PANC = sd(subset(epi_data, epi_data$T2D.PANC == 1)$AgeBaseline, na.rm = T)

# Sex distribution for all participants with a PANC diagnosis
female_count_T2D_PANC = sum(subset(epi_data, epi_data$T2D.PANC ==1)$Sex == 0)
male_count_T2D_PANC = sum(subset(epi_data, epi_data$T2D.PANC ==1)$Sex == 1)

perc_women_T2D_PANC = sum(subset(epi_data, epi_data$T2D.PANC ==1)$Sex == 0)/nrow(subset(epi_data, epi_da
perc_men_T2D_PANC = sum(subset(epi_data, epi_data$T2D.PANC ==1)$Sex == 1)/nrow(subset(epi_data, epi_data
```

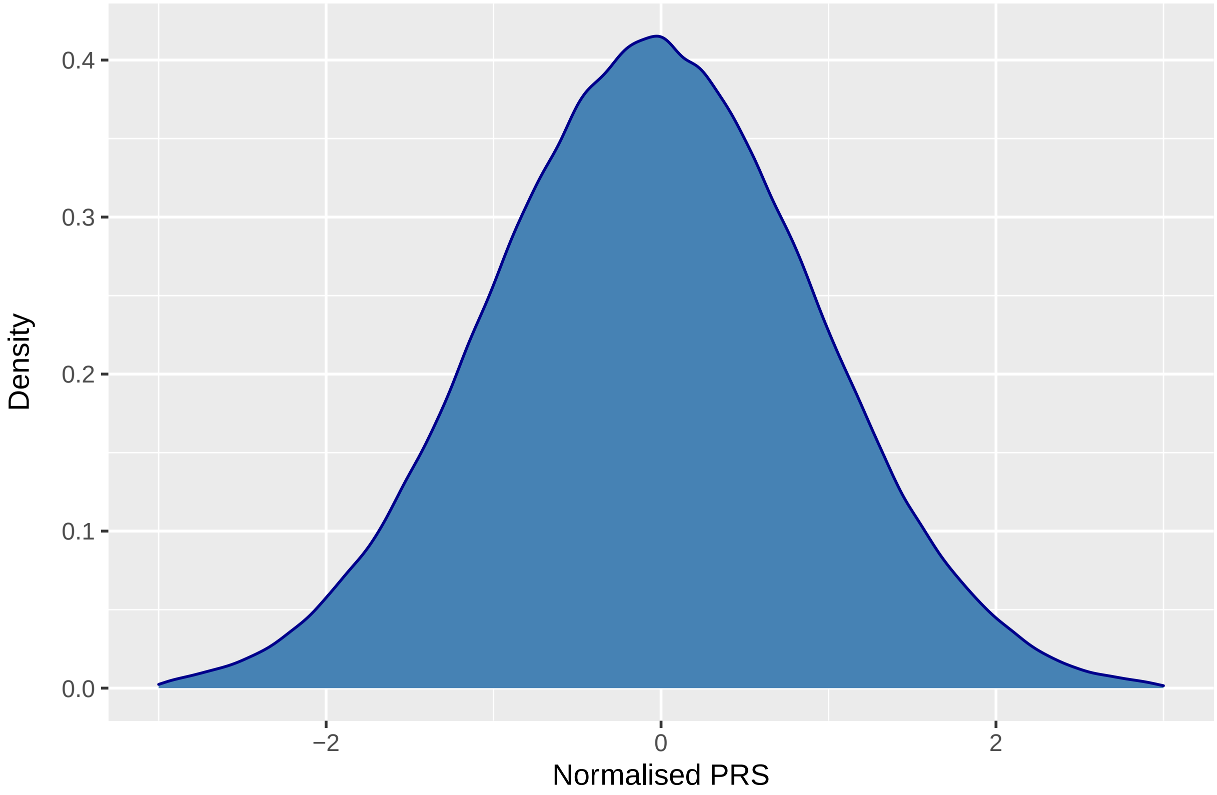
Polygenic Risk Scores (PRS)

Polygenic Risk Scores (PRS) are used to quantify an individual's genetic disease risk, by constructing a sum which is unweighted, or weighted by effect size, of all disease-associated risk alleles present in the individual's genome.

PRS for 1) T2D and 2) PanC have been calculated for each participant in the study population. The below plots show the distribution of the two PRS for the whole cohort. As expected, the PRS are normally distributed.



PANC PRS Distribution



T2D PRS Distribution