Клинико-лабораторная диагностика, ДЗ №1

Мироненко Ольга

```
library(tidyverse)
library(pROC)
knitr::opts_chunk$set(echo = TRUE, warning = FALSE,
                      message = FALSE, error = FALSE)

df <- read.csv("diabetes.csv")
summary(df)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin          BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780          Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437          1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725          Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719          Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262          3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200          Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

Задание 1

```
df <- df %>%
  # Заменим нули на пропускм для тех количественных переменных,
  # которые не могут принимать нулевые значения
  mutate_at(vars(Glucose, BloodPressure, SkinThickness, Insulin, BMI),
            ~ ifelse(. == 0, NA, .)) %>%
  # Переведем мг/дл в ммоль/л для глюкозы и выделим пациентов с НТГ
```

```
  mutate(Glucose_mml = Glucose/18,
         IGT = Glucose_mml >= 7.8)

table(df$IGT)
```

```
##
## FALSE  TRUE
##   571   192
```

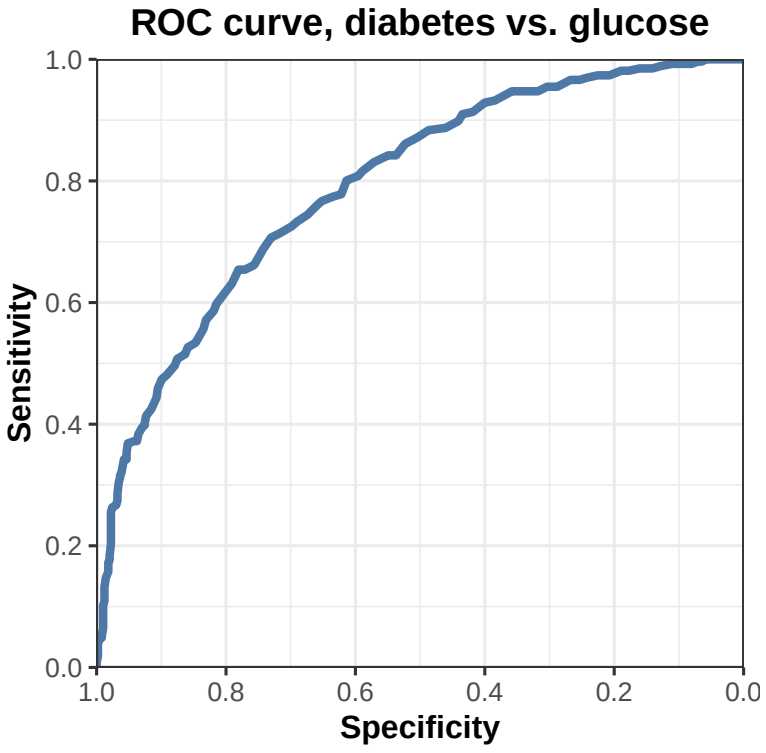
Таким образом, НТГ отсутствует у 571 пациента.

Задание 2

Построим ROC-кривую для предсказания сахарного диабета по уровню глюкозы, измеренному в ммоль/л.

```
roc_gluc <- roc(Outcome ~ Glucose_mml, data = df, ci = TRUE)

ggroc(roc_gluc, color = "#4E79A7", size = 1.5) +
  scale_x_reverse(expand = c(0,0),
                  breaks = seq(0,1,0.2)) +
  scale_y_continuous(expand = c(0,0),
                     breaks = seq(0,1,0.2)) +
  labs(x = "Specificity", y = "Sensitivity",
       title = "ROC curve, diabetes vs. glucose") +
  theme_bw(base_size = 12) +
  theme(axis.title = element_text(face = "bold"),
        plot.title = element_text(face = "bold", hjust = 0.5))
```



Задание 3

```
roc_gluc$auc
```

```
## Area under the curve: 0.7928
```

Задание 4

```
roc_gluc$ci
```

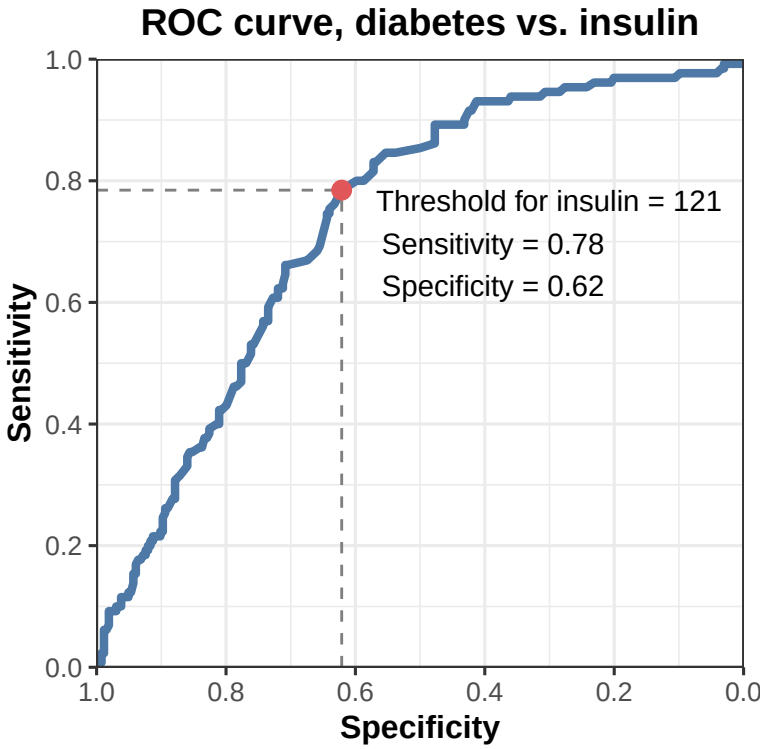
```
## 95% CI: 0.7599-0.8257 (DeLong)
```

Задание 5

Ответы на графике (в качестве оптимальной была выбрана точка, наиболее близкая к верхнему левому углу графика - я проверила, она совпадает с лучшим рещением по критерию Youden's J):

```
roc_ins <- roc(Outcome ~ Insulin, data = df, ci = TRUE)
roc_ins_best <- roc_ins %>%
  coords(x = "best", best.method = "closest.topleft") %>%
  mutate(lbl = sprintf("Threshold for insulin = %d\n  Sensitivity = %.2f\n
                       threshold, sensitivity, specificity))

ggroc(roc_ins, color = "#4E79A7", size = 1.5) +
  geom_line(aes(x = x, y = y),
            data.frame(x = c(1, roc_ins_best$specificity),
                       y = roc_ins_best$sensitivity),
            linetype = "dashed", size = 0.5, color =  "grey50") +
  geom_line(aes(x = x, y = y),
            data.frame(x = roc_ins_best$specificity,
                       y = c(0, roc_ins_best$sensitivity)),
            linetype = "dashed", size = 0.5, color =  "grey50") +
  geom_point(aes(x = specificity, y = sensitivity), roc_ins_best,
             color = "#E15759", size = 3) +
  geom_text(aes(x = specificity, y = sensitivity, label = lbl),
            roc_ins_best, hjust = -0.1, vjust = 1) +
  scale_x_reverse(expand = c(0,0),
                  breaks = seq(0,1,0.2)) +
  scale_y_continuous(expand = c(0,0),
                     breaks = seq(0,1,0.2)) +
  labs(x = "Specificity", y = "Sensitivity",
       title = "ROC curve, diabetes vs. insulin") +
  theme_bw(base_size = 12) +
  theme(axis.title = element_text(face = "bold"),
        plot.title = element_text(face = "bold", hjust = 0.5))
```



Задание 6

```
df %>%
  select(-Glucose, -IGT) %>%
  pivot_longer(cols = -Outcome, names_to = "Variable") %>%
  group_by(Variable) %>%
  summarise(AUC = roc(Outcome, value, ci = T)$ci[2] %>% round(3)) %>%
  arrange(-AUC) %>%
  gt::gt()
```

Variable	AUC
Glucose_mml	0.793
Insulin	0.732
Age	0.687
BMI	0.687
SkinThickness	0.663
Pregnancies	0.620
BloodPressure	0.608
DiabetesPedigreeFunction	0.606

Таким образом, максимальную площадь под ROC-кривой даёт предсказание диабета по уровню глюкозы: если исходить из определения, данного вами для сахарного диабета (хроническое эндокринное заболевание, сопровождающееся повышенным уровнем глюкозы), то это было ожидаемо, а точность прогнозирования по данному критерию не 100%-ная, скорее всего, потому, что это не единственный критерий постановки диагноза ``сахарный диабет", или потому, что нам неизвестно, как в этом датасете соотносятся моменты замеров глюкозы и статуса по диабету.

Минимальную площадь под кривой имеет вероятность наличия диабета на основании наследственного анамнеза - скорее всего, потому, что не все типы диабета являются наследственными/ развиваются вследствие генетической предрасположенности.