Predicting Readmission within 30 days for Diabetic Patients

B229592

December 11 2024

```
#loading libraries
library(sparklyr)
library(dplyr)
library(ggplot2)
library(knitr)
library(broom)
library(gt)
```

Introduction

Diabetes is a chronic disease of high blood sugar levels. It occurs naturally when the pancreas does not produce enough insulin, or it can be developed when production becomes reduced or the body gains resistance to insulin's effect.

The data is of patients admitted to hospital with diagnosed diabetes and their 30-day readmission rates. Data is taken from 130 US hospitals between 1999-2008. This analysis will predict readmissions using the clinical information recorded.

```
#creating spark connection from local device
sc <-  spark_connect(master = "local")
#importing csv using command that would work in big data setting
diabetes_spk <- spark_read_csv(sc, "diabetic_data.csv", overwrite = TRUE)

cached_diabetes_spk <- diabetes_spk %>%
  #filter out ">30" days, question asks within 30 days
  filter(readmitted != ">30")
```

Exploratory Data Analysis

There are nearly 50 different variables available to analyse. Variables such as race have not been analysed due to their arbitrary relationship to readmission rates. Variables such as weight and medical specialty have not been analysed as they have too many missing values to create valuable models from.

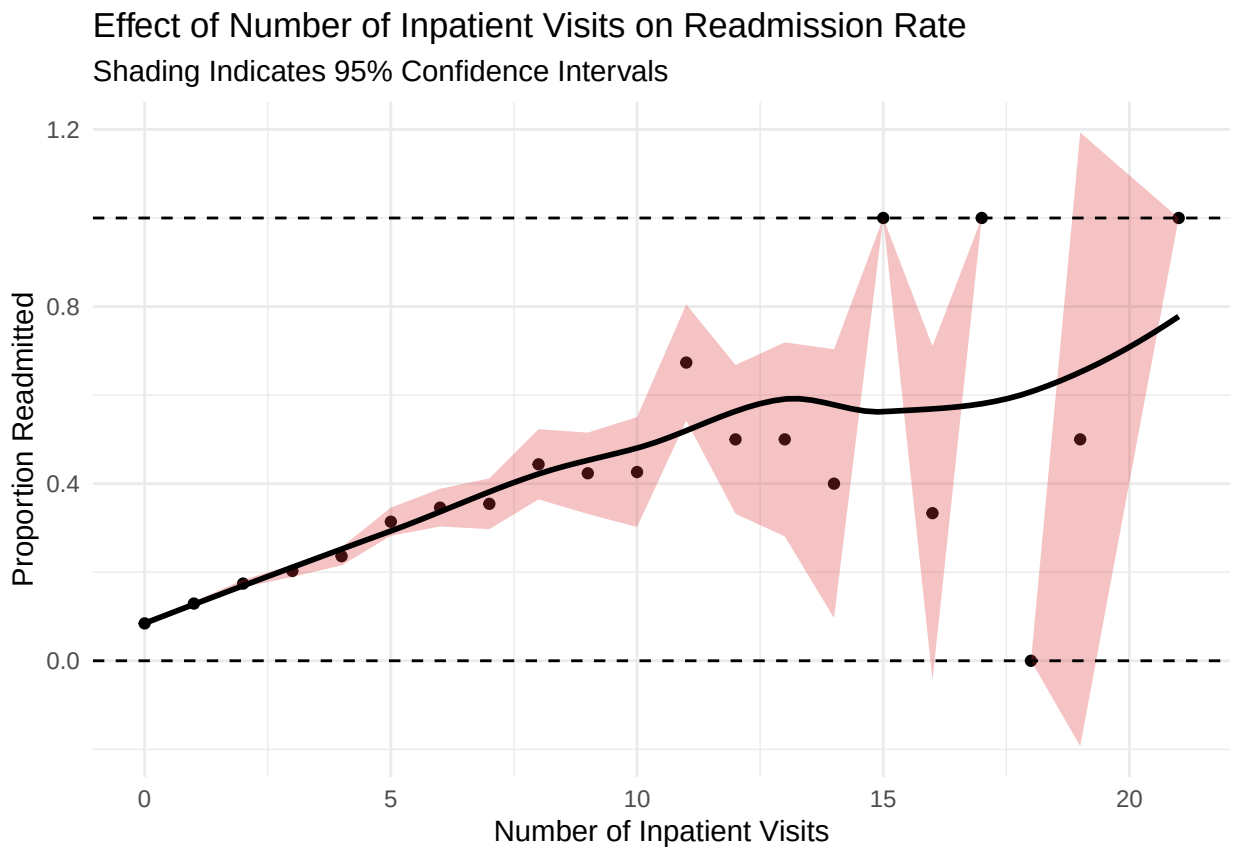
Effect of Number of Inpatient Visits

As patients readmitted must have gone into hospital previously, exploratory analysis will begin the number of admissions in the previous year. Only inpatient admissions are assessed as outpatient visits are likely less associated with chronic illness, such as breaking a finger, and are therefore less likely to affect readmission rates.

Figure 1

```
inpatient_visits <- diabetes_spk %>%
#calculating proportion of patients readmitted per group
  group_by(number_inpatient) %>%
    summarise(
    visit_count = n(),
    prop_readmitted = sum(ifelse(readmitted == "<30", 1, 0))/
                      n()
  ) %>%
#calculating confidence intervals for proportions
  mutate(std_error = sqrt(prop_readmitted * (1 - prop_readmitted) / visit_count),
         lower_ci = prop_readmitted - 1.96 * std_error,
         upper_ci = prop_readmitted + 1.96 * std_error,
         #ensuring treatment as number not factor
         number_inpatient = as.numeric(number_inpatient)
  ) %>%
  collect()


inpatient_visits %>%
  ggplot(aes(x = number_inpatient, y = prop_readmitted)) +
  #sindividual points
  geom_point() +
  # 95% onfidence intervals
  geom_ribbon(aes(
    ymin = lower_ci,
    ymax = upper_ci
  ), fill = "#DF3636", alpha = 0.3) +
  #trend line
  geom_smooth(se = FALSE, method = "loess", colour = "black") +
  labs(
    x = "Number of Inpatient Visits",
    y = "Proportion Readmitted",
    title = "Effect of Number of Inpatient Visits on Readmission Rate",
    subtitle = "Shading Indicates 95% Confidence Intervals"
  ) +
  theme_minimal() +
  #showing limits as proportion can only be between 0 and 1
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_hline(yintercept = 0, linetype = "dashed")
```



The graph shows a positive relationship between number of previous visits and readmission. The relation is proportional with small confidence intervals up to 10 visits, and then becomes more sporadic with larger confidence intervals, likely due to small sample sizes. Despite this ambiguity, the relationship is strong enough to qualify for assessment in the model.

Effect of Length of Stay on Readmission

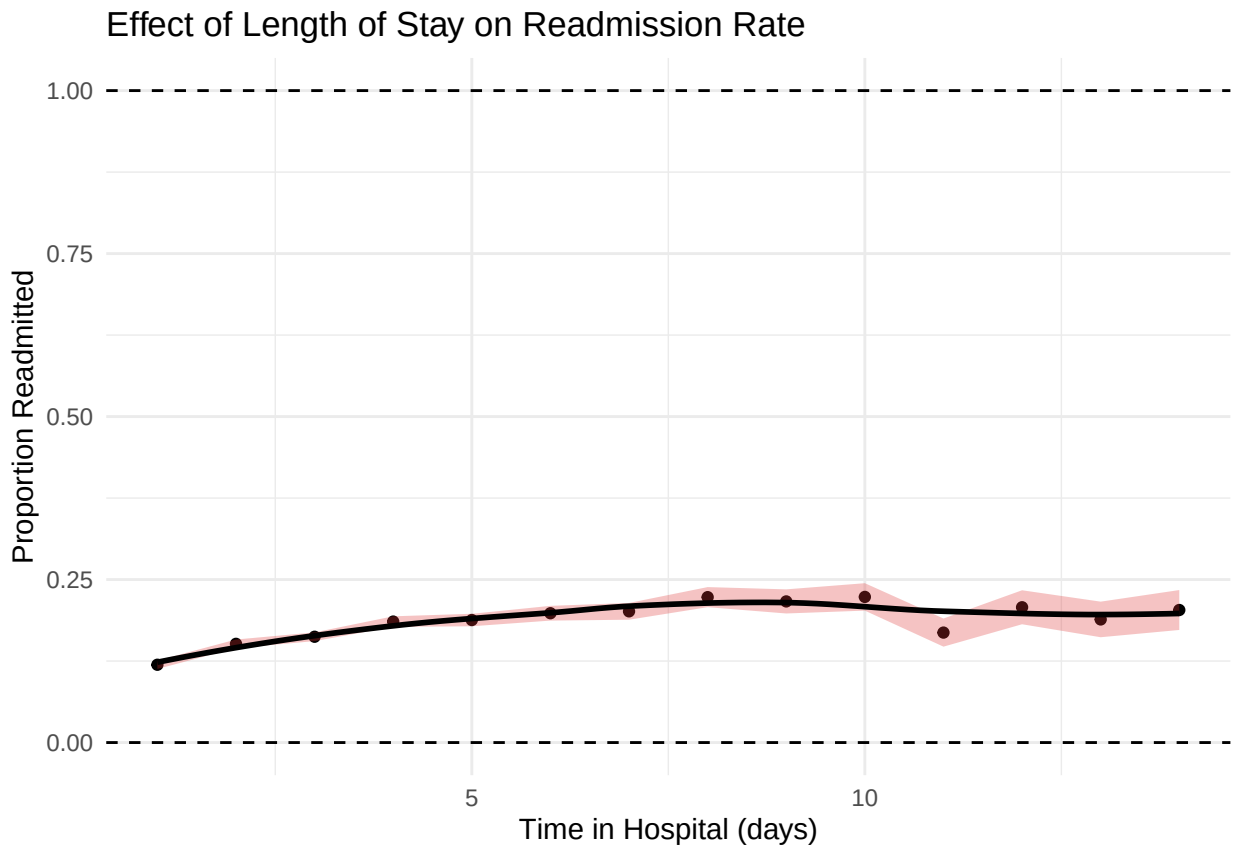
The duration of time spent in hospital is also likely a good indication of readmission due to the association with illness severity.

Figure 2

```
length_stay_plot <- cached_diabetes_spk %>%
  group_by(time_in_hospital) %>%
  #calculating proportion of readissions for each group
  summarise(
    length_count = n(),
    prop_readmitted = sum(ifelse(readmitted == "<30", 1, 0)) /
                      n()
  ) %>%
  mutate(std_error = sqrt(prop_readmitted * (1 - prop_readmitted) / length_count),
         lower_ci = prop_readmitted - 1.96 * std_error,
         upper_ci = prop_readmitted + 1.96 * std_error,
  ) %>%
```

```
  collect()

length_stay_plot %>%
    ggplot(aes(x = time_in_hospital, y = prop_readmitted)) +
  #individual points
  geom_point() +
  # 95% confidence intervals
  geom_ribbon(aes(
    ymin = lower_ci,
    ymax = upper_ci
  ), fill = "#DF3636", alpha = 0.3) +
  #trend line
  geom_smooth(se = FALSE, method = "loess", colour = "black") +
  labs(
    title = "Effect of Length of Stay on Readmission Rate",
    x = "Time in Hospital (days)",
    y = "Proportion Readmitted"
  ) +
  theme_minimal() +
  #showing limits as proportion can only be between 0 and 1
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_hline(yintercept = 0, linetype = "dashed")
```



Time in hospital also has a general positive relationship to readmission rates and is therefore will be evaluated in the model.

Effect of Specific Medications on Readmission

In the dataset, there are 24 medications. For each, patients are assigned to *No* if they were not on the medication, *Steady* if they were on the medication and their dose was unchanged, *Up* if they were on the medication and their dose was increased (Up), or *Down* if they were on the medication and their dose was decreased.

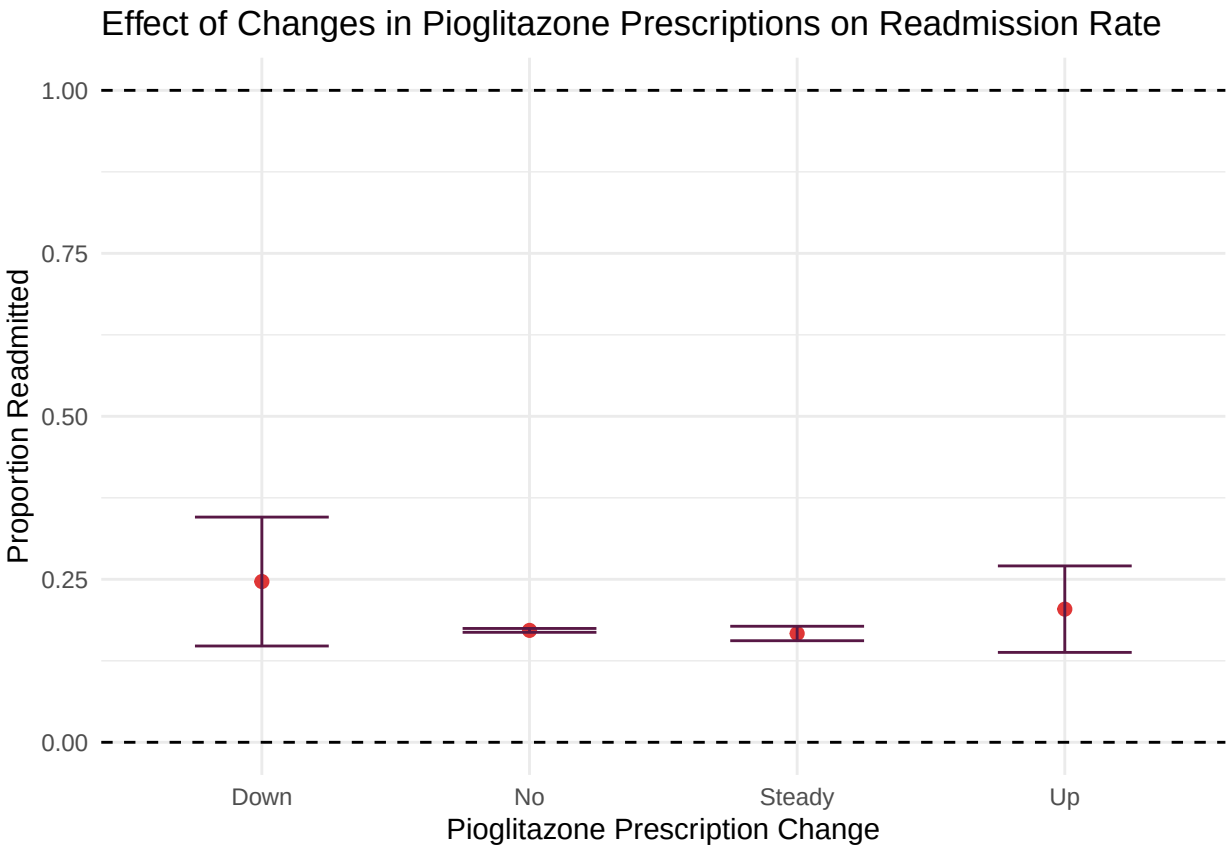
The common trend across all medications was that those with a decrease in prescription had the highest levels of readmission. This difference was most pronounced for miglitol and acarbose, but these medications had limited datapoints and large variation. Pioglitazone will be assessed as it has sufficient data with reasonable variance.

Figure 3

```
drug_graph <- cached_diabetes_spk %>%
  group_by(pioglitazone) %>%
    #calculating proportion of readissions for each group
  summarise(
    drug_count = n(),
    prop_readmitted = sum(ifelse(readmitted == "<30", 1, 0))/
                      n()
  ) %>%
    mutate(std_error = sqrt(prop_readmitted * (1 - prop_readmitted) / drug_count)
  ) %>%
  collect()

drug_graph %>%
  ggplot(aes(x = pioglitazone, y = prop_readmitted)) +
  #plotting proportion per group
  geom_point(size = 2, colour = "#DF3636") +
  #plotting 95% confidence interval
  geom_errorbar(aes( ymin = prop_readmitted - 1.96 * std_error,
                     ymax = prop_readmitted + 1.96 * std_error,),
                width = 0.5, colour = "#581845") +

  theme_minimal() +
  labs(
    title = "Effect of Changes in Pioglitazone Prescriptions on Readmission Rate",
    x = "Pioglitazone Prescription Change",
    y = "Proportion Readmitted"
  ) +
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_hline(yintercept = 0, linetype = "dashed")
```



Effect of Age on Readmissions

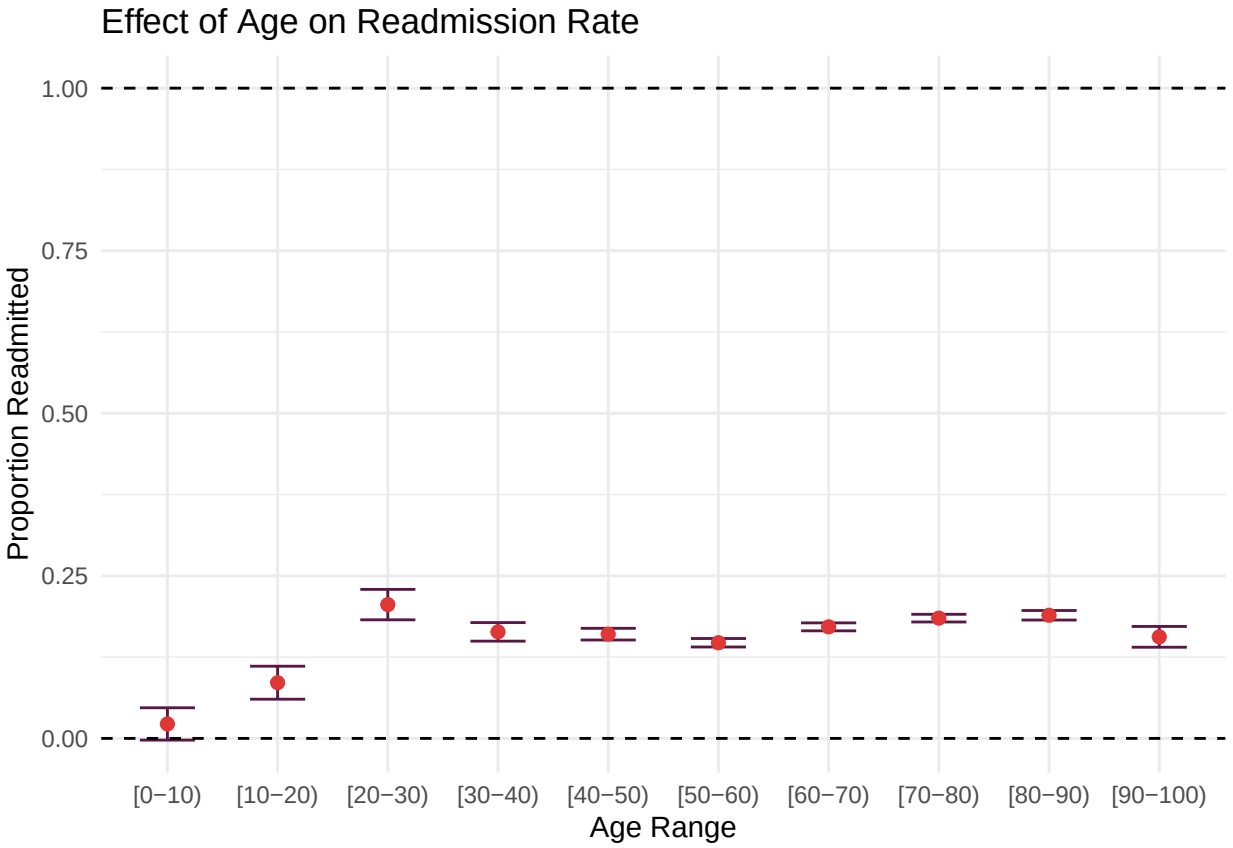
Younger populations are likely to be healthier due to better lifestyle (eg. exercise, diet, smoking) and could therefore be less likely to have complications leading to fewer readmissions.

Figure 4

```
age_graph <- cached_diabetes_spk %>%
  group_by(age) %>%
  summarise(
    age_count = n(),
    prop_readmitted = sum(ifelse(readmitted == "<30", 1, 0))/
                      n()
  ) %>%
    mutate(std_error = sqrt(prop_readmitted * (1 - prop_readmitted) / age_count)
  ) %>%
  collect()

age_graph %>%
  ggplot(aes(x = age, y = prop_readmitted)) +
  geom_errorbar(aes( ymin = prop_readmitted - 1.96 * std_error,
                     ymax = prop_readmitted + 1.96 * std_error,),
                width = 0.5, colour = "#581845") +
```

```
  geom_point(size = 2, colour = "#DF3636") +
theme_minimal() +
labs(
  title = "Effect of Age on Readmission Rate",
  x = "Age Range",
  y = "Proportion Readmitted"
)  +
geom_hline(yintercept = 1, linetype = "dashed") +
geom_hline(yintercept = 0, linetype = "dashed")
```



Age appears to significantly affect readmission. Patients aged between 0 and 20 have the lowest readmission rates, as hypothesised. Those between 20 and 30 appear to have the highest rates. This relationship will therefore be evaluated further in the model.

Modelling

The model used will be *binary logistic regression* as readmission is binary, using supervised learning to assess the model's suitability in a 'real' context.

There are certainly more factors affecting admission rates, but only 4 will be assessed to minimise chances of overfitting; **number of previous visits, length of stay, change in pioglitazone prescription and age**. There are no missing values for these variables in the dataset.

```r
model_diabetes_spk <- cached_diabetes_spk %>%
  mutate(
  # one hot encoding pioglitazone
  drug_steady = ifelse(pioglitazone == 'Steady', 1, 0),
  drug_none = ifelse(pioglitazone == 'No', 1, 0),
  drug_up = ifelse(pioglitazone == 'Up', 1, 0),
  drug_down = ifelse(pioglitazone == 'Down', 1, 0),
  # one hot encoding age
  ninety_onehundred = ifelse(age == '[90-100)', 1, 0),
  eighty_ninety = ifelse(age == '[80-90)', 1, 0),
  seventy_eighty = ifelse(age == '[70-80)', 1, 0),
  sixty_seventy = ifelse(age == '[60-70)', 1, 0),
  fifty_sixty = ifelse(age == '[50-60)', 1, 0),
  forty_fifty = ifelse(age == '[40-50)', 1, 0),
  thirty_forty = ifelse(age == '[30-40)', 1, 0),
  twenty_thirty = ifelse(age == '[20-30)', 1, 0),
  ten_twenty = ifelse(age == '[10-20)', 1, 0),
  zero_ten = ifelse(age == '[0-10)', 1, 0),
  # making readmissions binary
  readmitted_binary = ifelse(readmitted == 'NO', 0, 1)) %>%
  #splitting data into 80% training and 20% testing
  #seed ensures same set is re-run every analysis for reproducibility
  sdf_random_split(training = 0.8, test = 0.2, seed = 10)

# modelling using binary logistic regression
readmissions_model <-  ml_generalized_linear_regression(model_diabetes_spk$training,
                  readmitted_binary ~
                    number_inpatient +
                    time_in_hospital +
# drug_none is omitted as reference as it has the lowest proportion of readmissions
                    drug_up +
                    drug_down +
                    drug_steady +
# zero_ten is omitted as reference as it has the lowest proportion of readmissions
                    ninety_onehundred +
                    eighty_ninety +
                    seventy_eighty +
                    sixty_seventy +
                    fifty_sixty +
                    forty_fifty +
                    thirty_forty +
                    twenty_thirty +
                    ten_twenty,
#ensuring binomial logistic regression
      family = "binomial")
```

Results

```r
#tidying results
readmissions_results_spk <- tidy(readmissions_model)
```

Readmission Model Results

Model: Binary Logistic Regression

Predictor Variable	Odds Ratio	95% Confidence Interval
(Intercept)	0.025	0.01 - 0.08
number_inpatient	1.624	1.59 - 1.66
time_in_hospital	1.036	1.03 - 1.04
drug_up	1.537	0.98 - 2.41
drug_down	1.190	0.62 - 2.29
drug_steady	1.045	0.95 - 1.15
ninety_onehundred	4.533	1.43 - 14.41
eighty_ninety	5.567	1.76 - 17.58
seventy_eighty	5.431	1.72 - 17.14
sixty_seventy	5.207	1.65 - 16.44
fifty_sixty	4.253	1.35 - 13.43
forty_fifty	4.542	1.44 - 14.36
thirty_forty	4.500	1.42 - 14.28
twenty_thirty	4.486	1.4 - 14.36
ten_twenty	2.708	0.81 - 9.03

```
results_table <- readmissions_results_spk %>%
  mutate(
    #calculating Odds Ratios and their 95% confidence intervals
    OR = exp(estimate),
    OR_lower_confidence = exp(estimate - 1.96 * std.error),
    OR_upper_confidence = exp(estimate + 1.96 * std.error)
  ) %>%
  mutate(
    #showing the confidence interval as a range (character)
    OR_CI = paste0(round(OR_lower_confidence, 2), " - ", round(OR_upper_confidence, 2))
  ) %>%
  select(term, OR, OR_CI) %>%
  collect()

#displaying results in table
results_table %>%
  gt() %>%
  tab_header(
    title = "Readmission Model Results",
    subtitle = "Model: Binary Logistic Regression"
    ) %>%
  cols_align(columns = c(OR, OR_CI),
             align = "center") %>%
  fmt_number(columns = OR,
             decimals = 3) %>%
  cols_label(term = md("**Predictor Variable**"),
             OR = md("**Odds Ratio**"),
             OR_CI = md("**95% Confidence Interval**"))
```

Main Findings

Effect of Number of Inpatient Visits As the Odds Ratio and its 95% Confidence Interval are sufficiently above 1, we can conclude that the number of inpatient visits is associated with a significant increase in readmission rates, with each admission increasing chances of readmission by 62%.

Effect of Time in Hospital The Odds Ratio and its 95% Confidence Interval are also above 1, but only slightly. Therefore, we can conclude that time in hospital is associated with a slight but significant increase in readmission rates, with each day increasing chances by 3.6%.

Effect of Pioglitazone Prescription Changes The reference group is those with no pioglitazone prescribed. All other groups have an Odds Ratio above 1 meaning that there is an increased likelihood of readmission, but as the Confidence Intervals span over 1, this is not a statistically significant effect. The highest increase appears to be due to an increase in prescription dosage, rather than a decrease as initially suspected.

Effect of Age As specified in the model, the reference group is patients aged between 0 and 10. All other age groups have a higher chance of readmission shown by the Odds Ratios. Ages 10 to 20 have the lowest difference, although the Odds Ratio of readmission is nearly treble that of the reference group, the 95% Confidence Interval is wide, spanning over 1, indicating statistical insignificance but trending towards increased admissions. All other age groups show clear statistical significance with increasing readmission rates compared to those between 0 and 10 years old, but the extent of this difference cannot be conclusively determined due to the wide Confidence Intervals seen throughout.

Evaluation of Model

```
#predicting with test dataset
predictions_spk <- ml_predict(readmissions_model, model_diabetes_spk$test)

#using predictions to calculate confusion matrix counts
confusion_mx_calcs <- predictions_spk %>%
  mutate(
    accuracy = case_when(
      readmitted_binary == 0 & prediction < 0.5 ~ "TrueNeg",
      readmitted_binary == 0 & prediction >= 0.5 ~ "FalsePos",
      readmitted_binary == 1 & prediction < 0.5 ~ "FalseNeg",
      readmitted_binary == 1 & prediction >= 0.5 ~ "TruePos"
    )
  ) %>%
  summarise(
    TruePos = sum(ifelse(accuracy == "TruePos", 1, 0)),
    FalseNeg = sum(ifelse(accuracy == "FalseNeg", 1, 0)),
    TrueNeg = sum(ifelse(accuracy == "TrueNeg", 1, 0)),
    FalsePos = sum(ifelse(accuracy == "FalsePos", 1, 0))
  ) %>%
  mutate(
# calculating accuracy
    Accuracy = (TruePos + TrueNeg) /
               (TruePos + TrueNeg + FalsePos + FalseNeg)
  ) %>%
```

Model Evaluation Metrics

Model: Binary Logistic Regression

Metric	Value
Accuracy	0.835
PR-AUC	0.344
ROC-AUC	0.679

```
  collect()

#predicting ROC-AUC
AUC <- ml_binary_classification_evaluator(
  predictions_spk,
  label_col = "readmitted_binary",
  raw_prediction_col = "prediction",
  metric_name = "areaUnderROC"
  )

#predicting PR-AUC
AUPR <-  ml_binary_classification_evaluator(
  predictions_spk,
  label_col = "readmitted_binary",
  raw_prediction_col = "prediction",
  metric_name = "areaUnderPR"
  )

#creating dataframe combining metrics chosen
table <- tibble(
  Metric = c("Accuracy",
             "PR-AUC",
             "ROC-AUC"),
  Value = c(confusion_mx_calcs$Accuracy,
            AUPR,
            AUC)
)

#displaying table
table %>%
  gt() %>%
  tab_header(
    title = "Model Evaluation Metrics",
    subtitle = "Model: Binary Logistic Regression"
    ) %>%
  cols_align(columns = Value,
             align = "center") %>%
  fmt_number(columns = Value,
             decimals = 3) %>%
  cols_label(Metric = md("**Metric**"),
             Value = md("**Value**"))
```

Models with an accuracy of over 70% are considered a good model. The accuracy of this model is 83.5%

indicating very good performance. However, a good AUC statistic is considered between 0.7 to 0.9. The model's AUC-ROC is 0.68 indicating that our model is slightly under-performing and the AUC-PR is 0.34, indicating serious underperformance. This is likely due to the imbalance of readmission, and due to the much higher number of people not readmitted, the model is presumed to fit well.

Limitations

Dataset Limitations As the model uses data from 1999 to 2008, it will likely not be effective using contemporary data to differences in healthcare practices. For example, pioglitazone has been shown to increase bladder cancer incidence leading to reduced prescription in the US (Singh and Correa, 2020).

Additionally, the data is imbalanced in the outcome of readmissions, therefore future investigations should use more balanced samples or use oversampling with algorithms such as the SMOTE method to address this issue.

Variable Limitations The variables accessible do not include co-morbidities, which have substantial impact on health outcomes. It would be advantageous to be include these in the model.

Model Limitations Assessing 4 variables could be oversimplifying the problem, hence potential underfitting of readmission predictions. Additionally, exploratory analysis shows non-linear relationships which could indicate that binary logistic regression is not the most suitable for the model. Further analysis should include alternative models, such as random forests, and these should be compared using metrics such as Accuracy and AIC to assess which one is best.

```
spark_disconnect(sc)
```

References

Singh, G. and Correa, R. (2020). Pioglitazone. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK544287/.