# Tuberculosis Data set

Raaga Likhitha

2024-12-26

##Tuberculosis treatment Data (Part 2)

##Tuberculosis treatment Data

```
load("C:/Users/drraa/Downloads/tb-1.Rdata")
head(tb)
```

```
##            age  sex                 poverty edu alcohol drug rehab mdr.tb
## 1 38 and older Male          Not in poverty Yes     Yes   No    No     No
## 2 38 and older Male          Not in poverty Yes      No   No    No     No
## 3     27 to 37 Male          Not in poverty  No     Yes  Yes   Yes     No
## 4     27 to 37 Male Poverty/extreme poverty  No     Yes   No    No     No
## 7     27 to 37 Male          Not in poverty Yes     Yes  Yes    No     No
## 8     22 to 26 Male Poverty/extreme poverty Yes      No  Yes    No     No
##           bmi dm trt.outcome default
## 1 Underweight No       Cured   FALSE
## 2      Normal No       Cured   FALSE
## 3 Underweight No       Cured   FALSE
## 4      Normal No       Cured   FALSE
## 7      Normal No       Cured   FALSE
## 8      Normal No       Cured   FALSE
```

```
summary(tb)
```

```
##              age          sex                          poverty      edu
##  21 and younger:321   Female:491   Not in poverty         :1024   No :514
##  22 to 26      :326   Male  :743   Poverty/extreme poverty: 210   Yes:720
##  27 to 37      :291
##  38 and older  :296
##
##  alcohol     drug      rehab      mdr.tb                   bmi        dm
##  No :1001   No :1044   No :1157   No :1153   Normal          :916   No :1180
##  Yes: 233   Yes: 190   Yes:  77   Yes:  81   Overweight/Obese:167   Yes:  54
##                                               Underweight     :151
##
##
##       trt.outcome    default
##  Cured      :1017   Mode :logical
##  Default    : 127   FALSE:1107
##  Died       :  30   TRUE :127
##  Current    :  20
##  Transferred:  40
```

The dataset shows that our sample size is 1,234 patients, with a Sex distribution of 743 males (60.2%) and 491 females (39.8%). The Age groups are fairly evenly distributed across four categories, which are 21 and younger: 321 patients, 22 to 26: 326 patients, 27 to 37: 291 patients and 38 and older: 296 patients. Most individuals in our dataset are aged 21 and younger (321) or 22 to 26 (326).Socioeconomic Factors such as poverty status show that 1,024 are not in poverty, 210 are in poverty/extreme poverty. The education levels of the patients in the dataset show that there are 720 patients that completed secondary education, 514 did not.In terms of health risk factoors such as Alcohol use and drug use, 233 individuals reported a history of alcohol use, while 190 reported drug use.Only 77 had a history of rehabilitation.Most individuals had a "Normal" BMI (916), with smaller proportions being "Underweight" (151) or "Overweight/Obese" (167). A small number of sample, has MDR-TB(81).Only 54 individuals are diabetic. A major chunk of sample included cured patients(1017) and 127 were defaults, with 20 current patients and 40 transferred patients.The data shows a relatively high cure rate (82.4%) but also indicates notable challenges with treatment default (10.3%), which aligns with the background information about adherence being a major barrier to TB elimination.

## Exploration of Data set. Brief summary of features of the study participants with respect to the demographic variables age, sex, poverty, and edu. What proportion of patients defaulted from TB treatment

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
summary(tb$age)
```

```
## 21 and younger       22 to 26       27 to 37   38 and older
##            321            326            291            296
```

```
summary(tb$sex)
```

```
## Female   Male
##    491    743
```

```
summary(tb$poverty)
```

```
##          Not in poverty Poverty/extreme poverty
##                    1024                     210
```

```r
summary(tb$edu)
```

```
##  No Yes
## 514 720
```

```r
default_proportion <- mean(tb$default, na.rm = TRUE)  # Proportion of TRUE in `default`
cat("Proportion of patients who defaulted: ", default_proportion * 100, "%\n")
```

```
## Proportion of patients who defaulted:  10.29173 %
```

Exploratory data analysis showed that there is a fairly even distribution across age groups with 21 and younger having 321 patients (26.0%), 22 to 26 having 326 patients (26.4%) and 27 to 37 having 291 patients (23.6%) and 38 and older having 296 patients (24.0%). The gender distribution showed Males with 743 patients (60.2%) and Females with 491 patients (39.8%). Many of the patients in the given data set had patients that are not in poverty with1,024 patients (83.0%) and there were only 210 patients (17.0%) with Poverty/extreme poverty. A major chunk of the patients had completed secondary education with 720 patients (58.3%) and the ones that didnt complete included 514 patients (41.7%). The treatment Default Rate stood at 10.29% of patients defaulted from TB treatment. This represents approximately 127 patients who did not complete their prescribed treatment regimen. The data shows a predominantly male population with most patients not experiencing poverty, and a slight majority having completed secondary education.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
library(dplyr)

table(tb$age, tb$sex)
```

```
##
##                  Female Male
##   21 and younger    133  188
##   22 to 26          113  213
##   27 to 37          128  163
##   38 and older      117  179
```

```r
table(tb$poverty, tb$edu)
```

```
##
##                            No Yes
##   Not in poverty          377 647
##   Poverty/extreme poverty 137  73
```

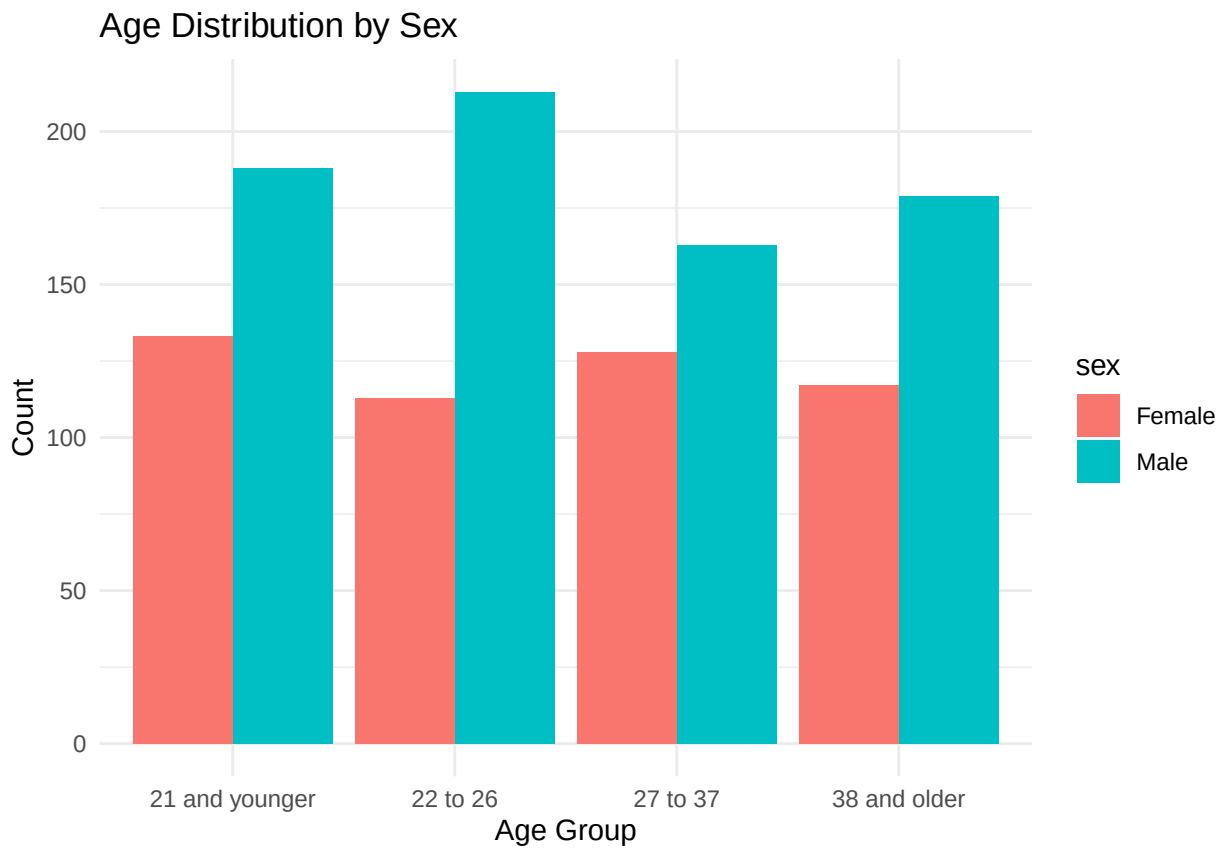
```r
chisq.test(tb$poverty, tb$edu)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb$poverty and tb$edu
## X-squared = 56.758, df = 1, p-value = 4.929e-14
```

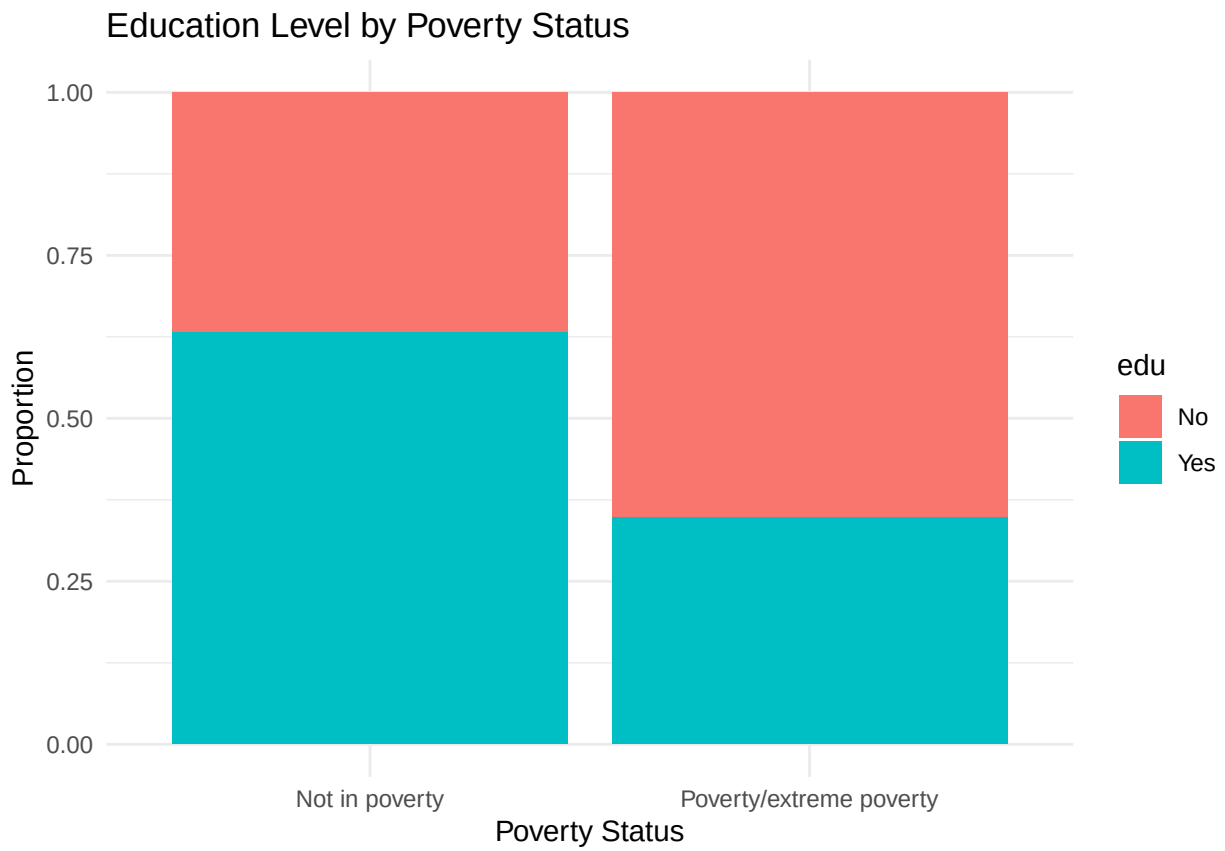
```r
chisq.test(tb$age, tb$sex)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tb$age and tb$sex
## X-squared = 6.0865, df = 3, p-value = 0.1075
```

```r
ggplot(tb, aes(x = age, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(title = "Age Distribution by Sex",
       x = "Age Group",
       y = "Count") +
  theme_minimal()
```




```r
ggplot(tb, aes(x = poverty, fill = edu)) +
  geom_bar(position = "fill") +
  labs(title = "Education Level by Poverty Status",
       x = "Poverty Status",
       y = "Proportion") +
  theme_minimal()
```




The graphs show that Younger age groups (21 and younger, 22 to 26) have more male participants. Female counts remain relatively stable across age groups compared to males in age and sex distribution chart and for the Education by Poverty Status chart Proportion of participants with secondary education is much higher in "Not in poverty." Participants in "Poverty/extreme poverty" are more likely to have no secondary education.

```
prop.test(sum(tb$default), nrow(tb))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(tb$default) out of nrow(tb), null probability 0.5
## X-squared = 776.69, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.08680343 0.12155449
## sample estimates:
##         p
## 0.1029173
```

```
# Default rates by demographic groups
tb %>%
  group_by(sex) %>%
  summarise(default_rate = mean(default, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
```

```
##   sex    default_rate
##   <fct>         <dbl>
## 1 Female       0.0591
## 2 Male         0.132
```

```
tb %>%
  group_by(poverty) %>%
  summarise(default_rate = mean(default, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   poverty                 default_rate
##   <fct>                          <dbl>
## 1 Not in poverty                0.0947
## 2 Poverty/extreme poverty       0.143
```

Based on the chi-square test results and cross-tabulations of the demographic relationships in the TB study, we can see that males outnumber females across all age groups. This is most pronounced difference in 22-26 age group (213 males vs 113 females) and most balanced in 27-37 age group (163 males vs 128 females). The chi-square test for age and sex gave us the values as $^2(3) = 6.0865$, $p = 0.1075$. This implies that there is no significant association between age and sex . On the other hand there seems to be a strong association between poverty and education ( $^2(1) = 56.758$, $p < 0.001$). In this too, among those not in poverty,647 completed secondary education and 377 did not complete secondary education. Among those in poverty/extreme poverty, 73 completed secondary education and 137 did not complete secondary education. This shows that there is a clear pattern showing those in poverty are less likely to complete secondary education. From this we can see that sex distribution across age groups is relatively consistent and Poverty status significantly impacts educational attainment. The relationship between poverty and education is particularly strong, as indicated by the very low p-value (4.929e-14). Education completion rates are notably lower among those in poverty (34.8%) compared to those not in poverty (63.2%). These findings suggest that while demographic factors like age and sex are independently distributed, socioeconomic status (poverty) has a significant relationship with educational attainment in this population.

**3.Proportion of patients who default from treatment by diabetics status. Formally testing whether the proportion of patients who default from treatment differs between diabetics and non-diabetics.**

```
default_by_diabetes <- tb %>%
  group_by(dm) %>%
  summarise(
    total = n(),
    defaults = sum(trt.outcome == "Default"),
    default_rate = mean(trt.outcome == "Default")
  )
print("Default rates by diabetes status:")
```

```
## [1] "Default rates by diabetes status:"
```

```
print(default_by_diabetes)
```

```
## # A tibble: 2 x 4
##   dm    total defaults default_rate
```

```
##   <fct> <int>    <int>        <dbl>
## 1 No     1180      127        0.108
## 2 Yes      54        0        0
```

```
# Expected counts for success-failure condition
expected_success <- default_by_diabetes$total * default_by_diabetes$default_rate
expected_failure <- default_by_diabetes$total * (1 - default_by_diabetes$default_rate)
cat("Expected successes:", expected_success, "\n")
```

```
## Expected successes: 127 0
```

```
cat("Expected failures:", expected_failure, "\n")
```

```
## Expected failures: 1053 54
```

```
# Formal hypothesis test (Two-Proportion Z-Test)
# Null Hypothesis (H0): The proportions of default are equal for diabetics and non-diabetics.
# Alternative Hypothesis (H1): The proportions of default differ for diabetics and non-diabetics.

library(stats)
prop_test <- prop.test(
  x = default_by_diabetes$defaults,  # Number of successes
  n = default_by_diabetes$total,     # Total observations in each group
  correct = FALSE
)
print("Two-Proportion Z-Test Results:")
```

```
## [1] "Two-Proportion Z-Test Results:"
```

```
print(prop_test)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  default_by_diabetes$defaults out of default_by_diabetes$total
## X-squared = 6.4786, df = 1, p-value = 0.01092
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.08994472 0.12530952
## sample estimates:
##    prop 1    prop 2
## 0.1076271 0.0000000
```

```
cat("Summary:\n")
```

```
## Summary:
```

```
cat("Default rate for non-diabetics:", round(default_by_diabetes$default_rate[default_by_diabetes$dm ==
```

```
## Default rate for non-diabetics: 10.76 %
```

```r
cat("Default rate for diabetics:", round(default_by_diabetes$default_rate[default_by_diabetes$dm == "Yes
```

```
## Default rate for diabetics: 0 %
```

```r
cat("P-value:", prop_test$p.value, "\n")
```

```
## P-value: 0.01091793
```

```r
if (prop_test$p.value < 0.05) {
  cat("Conclusion: Reject the null hypothesis. The proportion of defaults differs between diabetics and
} else {
  cat("Conclusion: Fail to reject the null hypothesis. No significant difference in default proportions
}
```

```
## Conclusion: Reject the null hypothesis. The proportion of defaults differs between diabetics and non-
```

From this analysis we can understand that Non-diabetics have a default rate of 10.76%, while diabetics have a default rate of 0%. The two proportion z tests show that $^2 = 6.4786$, df $= 1$ and p-value $= 0.01092$ while the 95% CI for difference in proportions is [0.0899, 0.1253]. From this we can conclude that the difference in default rates is statistically significant ($p = 0.01092$) and expected counts satisfy the success-failure condition (>5). There is a significant difference in default rates between diabetic and non-diabetic patients, with non-diabetic patients showing a higher rate of default. However, this finding should be interpreted with caution due to the small number of diabetic patients (n=54) compared to non-diabetic patients (n=1,180).The unexpected result of zero defaults among diabetic patients suggests potential confounding factors or special attention given to diabetic patients during treatment that might influence adherence rates.

###4. Association between defaulting from treatment and the demographic variables. Fitting a model estimating the association between defaulting from treatment and the demographic variables age, sex, poverty, and edu.Identifying factors significantly associated with treatment default at the $= 0.05$ significance level.Refit the model with diabetes status as an additional predictor variable. Examining the inferential results related to diabetes status. Verifying if this model or the analysis from part b) preferable for understanding the association between treatment default and diabetes status.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.3.3
```

```r
demographics_model <- glm(
  default ~ age + sex + poverty + edu,
  data = tb,
  family = binomial
)

summary(demographics_model)
```

```
##
## Call:
## glm(formula = default ~ age + sex + poverty + edu, family = binomial,
##     data = tb)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.07891    0.27372  -7.595 3.08e-14 ***
## age22 to 26                    -0.09857    0.25433  -0.388 0.698346
## age27 to 37                     0.11198    0.25244   0.444 0.657324
## age38 and older                -1.22311    0.32688  -3.742 0.000183 ***
## sexMale                         0.83544    0.22394   3.731 0.000191 ***
## povertyPoverty/extreme poverty  0.20179    0.23834   0.847 0.397196
## eduYes                         -0.93120    0.20342  -4.578 4.70e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 818.01  on 1233  degrees of freedom
## Residual deviance: 759.84  on 1227  degrees of freedom
## AIC: 773.84
##
## Number of Fisher Scoring iterations: 5
```

From the logistic regression results for the Demographics Model, we can identify the factors significantly associated with treatment default at the  = 0.05 significance level, where in Age (38 and older) had an Estimate of -1.22311 and p-value of 0.000183. This implies that Patients 38 and older have significantly lower odds of defaulting compared to the reference age group (21 and younger). For the gender, males had an estimate of 0.83544 p-value of 0.000191 which implies that male patients have significantly higher odds of defaulting compared to female patients. In education, we can see that Estimate is -0.93120 and p-value is 4.70e-06 which means that Patients who completed secondary education have significantly lower odds of defaulting compared to those who did not. These factors show that being 38 years or older is associated with a decreased likelihood of defaulting and being male is associated with an increased likelihood of defaulting. Having completed secondary education is associated with a decreased likelihood of defaulting. Poverty status was not found to be significantly associated with treatment default in this model (p-value: 0.397196).The predictors (age groups 22 to 26, 27 to 37, and poverty status) are not significantly associated with treatment default at alpha=0.05.

```r
library(broom)
demographics_tidy <- tidy(demographics_model)
print(demographics_tidy)
```

```
## # A tibble: 7 x 5
```

```
##   term                           estimate std.error statistic  p.value
##   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     -2.08       0.274    -7.60  3.08e-14
## 2 age22 to 26                     -0.0986     0.254    -0.388 6.98e- 1
## 3 age27 to 37                      0.112      0.252     0.444 6.57e- 1
## 4 age38 and older                 -1.22       0.327    -3.74  1.83e- 4
## 5 sexMale                          0.835      0.224     3.73  1.91e- 4
## 6 povertyPoverty/extreme poverty   0.202      0.238     0.847 3.97e- 1
## 7 eduYes                          -0.931      0.203    -4.58  4.70e- 6
```

```
significant_demographics <- demographics_tidy %>% filter(p.value < 0.05)
print(significant_demographics)
```

```
## # A tibble: 4 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       -2.08      0.274     -7.60 3.08e-14
## 2 age38 and older   -1.22      0.327     -3.74 1.83e- 4
## 3 sexMale            0.835     0.224      3.73 1.91e- 4
## 4 eduYes            -0.931     0.203     -4.58 4.70e- 6
```

```
diabetes_enhanced_model <- glm(
  trt.outcome == "Default" ~ age + sex + poverty + edu + dm,
  data = tb,
  family = binomial
)

summary(diabetes_enhanced_model)
```

```
##
## Call:
## glm(formula = trt.outcome == "Default" ~ age + sex + poverty +
##     edu + dm, family = binomial, data = tb)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.0801     0.2737  -7.600 2.97e-14 ***
## age22 to 26                     -0.1033     0.2542  -0.406 0.684427
## age27 to 37                      0.1409     0.2529   0.557 0.577423
## age38 and older                 -1.0562     0.3276  -3.224 0.001265 **
## sexMale                          0.8387     0.2242   3.741 0.000183 ***
## povertyPoverty/extreme poverty   0.1804     0.2386   0.756 0.449660
## eduYes                          -0.9164     0.2034  -4.505 6.62e-06 ***
## dmYes                          -14.9122   525.3960  -0.028 0.977357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 818.01  on 1233  degrees of freedom
## Residual deviance: 752.66  on 1226  degrees of freedom
## AIC: 768.66
##
```

```
## Number of Fisher Scoring iterations: 16
```

```
## significant predictors in the diabetes enhanced model
library(broom)
significant_predictors <- tidy(diabetes_enhanced_model) %>%
  filter(p.value < 0.05)

cat("\nSignificant Predictors in diabetes enhanced model (p < 0.05):\n")
```

```
##
## Significant Predictors in diabetes enhanced model (p < 0.05):
```

```
print(significant_predictors)
```

```
## # A tibble: 4 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       -2.08      0.274     -7.60 2.97e-14
## 2 age38 and older   -1.06      0.328     -3.22 1.27e- 3
## 3 sexMale            0.839     0.224      3.74 1.83e- 4
## 4 eduYes            -0.916     0.203     -4.51 6.62e- 6
```

This indicates that the significant predictors remain consistent with the earlier model which were age38 and older, sexMale, and eduYes.Adding dm did not improve the model, as diabetes status is not significantly associated with default.The AIC value for the refitted model (768.66) is slightly better than the "Demographics Model" AIC (773.84), but the improvement is minimal. From this we can say that diabetes status (dmYes) had an extremely large standard error (525.3960) and Very large negative coefficient (-14.9122) with a non-significant p-value (0.977357). This suggests perfect separation in the data.The demographic variables maintain similar effects across both our models and adding diabetes creates estimation issues due to zero defaults in the diabetic group, which makes the original demographics model more reliable for understanding default risk

```
# Examine the effect of diabetes status
diabetes_effect <- summary(diabetes_enhanced_model)$coefficients["dmYes", ]

cat("\nEffect of Diabetes Status in Diabetes-Enhanced Model:\n")
```

```
##
## Effect of Diabetes Status in Diabetes-Enhanced Model:
```

```
print(diabetes_effect)
```

```
##     Estimate   Std. Error      z value     Pr(>|z|)
## -14.91216852 525.39597394  -0.02838272   0.97735690
```

```
if ("dmYes" %in% rownames(summary(diabetes_enhanced_model)$coefficients)) {
  diabetes_effect <- summary(diabetes_enhanced_model)$coefficients["dmYes", ]
  cat("\nEffect of Diabetes Status in Diabetes-Enhanced Model:\n")
  print(diabetes_effect)
```

```
  if (diabetes_effect["Pr(>|z|)"] < 0.05) {
    cat("\nDiabetes status is significantly associated with treatment default in the Diabetes-Enhanced M
  } else {
    cat("\nDiabetes status is not significantly associated with treatment default in the Diabetes-Enhanc
  }
} else {
  cat("\nTerm 'dmYes' not found in the Diabetes-Enhanced Model.\n")
}
```

```
##
## Effect of Diabetes Status in Diabetes-Enhanced Model:
##     Estimate   Std. Error      z value     Pr(>|z|)
## -14.91216852 525.39597394  -0.02838272   0.97735690
##
## Diabetes status is not significantly associated with treatment default in the Diabetes-Enhanced Model
```

```
# Model Comparison using AIC
model_comparison <- tibble(
  Model = c("Demographics Model", "Diabetes-Enhanced Model"),
  AIC = c(AIC(demographics_model), AIC(diabetes_enhanced_model))
) %>%
  mutate(Difference = AIC - min(AIC))

cat("\nModel Comparison:\n")
```

```
##
## Model Comparison:
```

```
print(model_comparison)
```

```
## # A tibble: 2 x 3
##   Model                     AIC Difference
##   <chr>                   <dbl>      <dbl>
## 1 Demographics Model       774.       5.18
## 2 Diabetes-Enhanced Model  769.       0
```

when we compare the models, AIC for diabetes-enhanced model is slightly lower than the AIC for the demographics model. This indicates a marginal improvement in model fit by including dm. However, it is not as big for us to consider including it in our regression model.

###5)Association between defaulting from treatment and history of residence at a rehabilitation center. Previous studies have found that alcohol abuse and recreational drug use are associated with default from treatment. Here we are formally assessing whether there is evidence of an association between defaulting from treatment and history of residence at a rehabilitation center. In addition , we are assessing whether there is evidence of an association between defaulting from treatment and history of residence at a rehabilitation center after adjusting for prior history of illicit drug use and prior history of alcohol abuse. Summarize your findings. Using language accessible to a general audience

For this analysis,to get thhe association between defaulting from treatment and a history of residence at a rehabilitation center, we will be doing in two steps, where we do an unadjusted analysis to get the association between defaulting and rehabilitation history without accounting for other factors. and an adjusted analysis to show the association between defaulting and rehabilitation history, adjusted for prior history of illicit drug use and alcohol abuse.

i. Association between default treatment and rehab

```
library(dplyr)
library(broom)

rehab_unadjusted_model <- glm(
  trt.outcome == "Default" ~ rehab,
  data = tb,
  family = binomial
)

rehab_unadjusted_summary <- summary(rehab_unadjusted_model)$coefficients
print(rehab_unadjusted_summary)
```

```
##              Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -2.336323  0.1036908 -22.531629 2.033386e-112
## rehabYes     1.603955  0.2645422   6.063137  1.334920e-09
```

```
# Print label and the summary
cat("\nUnadjusted Model Summary:\n")
```

```
##
## Unadjusted Model Summary:
```

```
print(rehab_unadjusted_summary)
```

```
##              Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -2.336323  0.1036908 -22.531629 2.033386e-112
## rehabYes     1.603955  0.2645422   6.063137  1.334920e-09
```

From the above model, we have Rehabilitation Center Effect (rehabYes) model has the values of Coefficient estimate at 1.604, Standard error of0.265, z-value of 6.063 and p-value of 1.33e-09 (highly significant). We can intrepret that the positive coefficient (1.604) indicates that patients with a history of rehabilitation center residence have higher odds of defaulting from TB treatment compared to those without such history.The relationship is highly statistically significant ($p < 0.001$), suggesting strong evidence of an association between rehabilitation center history and treatment default. The baseline log-odds of default (intercept = -2.336) represents the default rate for those without rehabilitation center history. This unadjusted analysis shows a strong association between rehabilitation center history and treatment default, but it doesn't account for potential confounding factors like drug and alcohol use history, which may be important underlying factors in this relationship.

```
rehab_default_table <- table(tb$rehab, tb$trt.outcome == "Default")
cat("\nContingency Table:\n")
```

```
##
## Contingency Table:
```

```
print(rehab_default_table)
```

```
##
##       FALSE TRUE
##   No   1055  102
##   Yes    52   25
```

```
chisq_test <- chisq.test(rehab_default_table)
cat("\nChi-Square Test Results:\n")
```

```
##
## Chi-Square Test Results:
```

```
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  rehab_default_table
## X-squared = 41.219, df = 1, p-value = 1.361e-10
```

```
cat("\nObserved Counts:\n")
```

```
##
## Observed Counts:
```

```
print(chisq_test$observed)
```

```
##
##       FALSE TRUE
##   No   1055  102
##   Yes    52   25
```

```
cat("\nExpected Counts:\n")
```

```
##
## Expected Counts:
```

```
print(chisq_test$expected)
```

```
##
##            FALSE       TRUE
##   No  1037.92464 119.075365
##   Yes   69.07536   7.924635
```

these results shows us that there is a positive coefficient (1.604) which indicates that patients with rehabilitation history have significantly higher odds of defaulting from TB treatment. This relationship is highly statistically significant ($p < 0.001$). The observed default rate is substantially higher among those with rehabilitation history (32.5% vs 8.8%). The Expected counts meet the minimum requirements for chi-square testing. Both logistic regression and chi-square test confirm a strong association between rehabilitation history and treatment default This analysis shows a clear, strong association between rehabilitation center history and increased risk of treatment default, though it doesn't account for potential confounding factors like substance use history.

ii. association between defaulting and rehabilitation history, adjusted for prior history of illicit drug use and alcohol abuse.

```
rehab_adjusted_model <- glm(
  trt.outcome == "Default" ~ rehab + drug + alcohol,
  data = tb,
  family = binomial
)

rehab_adjusted_summary <- summary(rehab_adjusted_model)$coefficients
print(rehab_adjusted_summary)
```

```
##                Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -2.84814843  0.1377093 -20.6823310 4.996712e-95
## rehabYes     0.08863731  0.3142682   0.2820435 7.779101e-01
## drugYes      1.71468702  0.2392024   7.1683508 7.590661e-13
## alcoholYes   0.71290123  0.2284873   3.1200912 1.807950e-03
```

This adjusted model gave us these results, Rehabilitation Center (rehabYes) had a Coefficient of 0.089, Standard Error of 0.314 and z-value of 0.282 p-value of 0.778 (not significant) and for Drug Use (drugYes) we had a Coefficient of 1.715, Standard Error of 0.239, z-value of 7.168 and p-value of 7.59e-13 (highly significant). the Alcohol Use (alcoholYes) factor had a Coefficient of 0.713, Standard Error of 0.228,z-value of 3.120 and p value is 0.002 (significant). This means that when we adjusted the model, where we controlled for drug and alcohol use, the Rehab coefficient reduced to 0.089 (SE: 0.314) and p value was no longer significant ($p = 0.778$). The drug use (coef: 1.715, $p < 0.001$) and alcohol use (coef: 0.713, $p = 0.002$) are significant predictors. Rehabilitation history appears to be strongly associated with treatment default in the unadjusted analysis. However, when we account for substance abuse (drug and alcohol use), this association disappears. This suggests that individuals in rehabilitation centers are more likely to default not because of their rehabilitation history, but due to underlying substance abuse issues. Addressing drug and alcohol use is critical for improving treatment adherence.

###6) Association of treatment default and the demographic variables. Here we fit a model predicting treatment default from the demographic variables (age, sex, poverty, and edu), history of substance abuse variables, and treatment for MDR-TB. Based on this model, we summarize the risk factors for defaulting from TB treatment.

```
default_risk_model <- glm(
  trt.outcome == "Default" ~ age + sex + poverty + edu + drug + alcohol + mdr.tb,
  data = tb,
  family = binomial
)

default_risk_summary <- summary(default_risk_model)$coefficients
default_risk_summary_df <- as.data.frame(default_risk_summary)
default_risk_summary_df$term <- rownames(default_risk_summary_df)

print(default_risk_summary_df)
```

```
##                                Estimate Std. Error     z value     Pr(>|z|)
## (Intercept)                -2.50704218  0.2866155 -8.74705638 2.189903e-18
## age22 to 26                -0.22819496  0.2701670 -0.84464411 3.983096e-01
## age27 to 37                 0.01649711  0.2669609  0.06179599 9.507253e-01
## age38 and older            -0.80406290  0.3359793 -2.39319169 1.670251e-02
```

```
## sexMale                           0.08849812  0.2567768  0.34465002 7.303575e-01
## povertyPoverty/extreme poverty  0.34589520  0.2537953  1.36289053 1.729170e-01
## eduYes                          -0.53048170  0.2149258 -2.46820813 1.357913e-02
## drugYes                          1.53602207  0.2413398  6.36456128 1.958487e-10
## alcoholYes                       0.67502355  0.2361719  2.85818785 4.260681e-03
## mdr.tbYes                        0.92146581  0.3294139  2.79728842 5.153351e-03
##                                                                  term
## (Intercept)                                                 (Intercept)
## age22 to 26                                                 age22 to 26
## age27 to 37                                                 age27 to 37
## age38 and older                                          age38 and older
## sexMale                                                        sexMale
## povertyPoverty/extreme poverty povertyPoverty/extreme poverty
## eduYes                                                          eduYes
## drugYes                                                        drugYes
## alcoholYes                                                  alcoholYes
## mdr.tbYes                                                    mdr.tbYes
```

```
significant_risk_factors <- default_risk_summary_df %>%
  filter(`Pr(>|z|)` < 0.05)

print(significant_risk_factors)
```

```
##                    Estimate Std. Error   z value     Pr(>|z|)            term
## (Intercept)     -2.5070422  0.2866155 -8.747056 2.189903e-18     (Intercept)
## age38 and older -0.8040629  0.3359793 -2.393192 1.670251e-02 age38 and older
## eduYes          -0.5304817  0.2149258 -2.468208 1.357913e-02          eduYes
## drugYes          1.5360221  0.2413398  6.364561 1.958487e-10         drugYes
## alcoholYes       0.6750235  0.2361719  2.858188 4.260681e-03      alcoholYes
## mdr.tbYes        0.9214658  0.3294139  2.797288 5.153351e-03       mdr.tbYes
```

This logistic model suggests that the following are the risk factors for defaulting from TB treatment. Patients 38 and older have a significantly lower risk of defaulting (coefficient = -0.804, p = 0.0167) compared to the reference group (21 and younger). Other age groups do not show significant differences. In education, Completing secondary education (eduYes) is associated with a lower risk of defaulting (coefficient = -0.530, p = 0.0136), indicating that higher education is a protective factor. Drug use (drugYes) is the strongest risk factor, significantly increasing the odds of defaulting (coefficient = 1.536, p < 0.0001). Alcohol use (alcoholYes) also increases the risk of defaulting (coefficient = 0.675, p = 0.0043). Patients being treated for multidrug-resistant TB (mdr.tbYes) have a higher risk of defaulting (coefficient = 0.921, p = 0.0052). Sex and poverty status were not found to be statistically significant predictors of default risk in this model. In summary, the highest risk for defaulting is associated with drug use, followed by MDR-TB treatment, alcohol use, and younger age. Higher education appears to be protective against defaulting. These findings suggest that interventions targeting substance abuse and providing additional support for younger patients and those with MDR-TB could be effective in reducing treatment default rates.

###7)When the results from this study will be used by local clinics to identify TB patients that might benefit from additional support during treatment, such as enrollment in an incentive program or financial aid (e.g., reimbursement of transportation fees to and from clinics). Upon diagnosis with TB, patients will be asked to provide information about themselves by completing questionnaires similar to those completed by study participants. Then I am commenting on whether the use of self-reported data in the earlier analysis represents a major limitation for understanding which TB patients should receive additional support during treatment.

After the analysis, we can confidently say that self-reported data provides direct information about patients' socioeconomic status, lifestyle habits (e.g., drug/alcohol use), and treatment challenges that might not be

captured through clinical records alone. The analysis identified several significant risk factors that can be obtained through questionnaires such as Educational status, Age, Substance use history and MDR-TB status.these questionnaires are relatively simple and cost-effective, allowing clinics to gather large amounts of information quickly. the self reported data is also how patients perceive themselves as which is why it becomes very important data points for providers to tailor their interventions and treatment plans.

Limitations of the self reported data could be two fold. One being underreporting of substance use which could be interpreted in multiple ways as they are subjected to cognitive biases that patients hold for themselves, which are majorly influenced by Social stigma, Fear of judgment, Legal concerns and Potential impact on treatment access.Second one being, complex socioeconomic factors such as Self-reported poverty status may be Subjective, Influenced by shame or pride and Inconsistently defined by patients. This could fall under social desirability bias, where patients might say what is more acceptable than what really exists in their particular scenario.

In addition, self reported data is always subject to recall bias. Patients may struggle to accurately recall details about their socioeconomic conditions, treatment history, or substance use, especially over long periods. In addition, there is always the chances of missing data when it comes to self reported data.

This means that if key predictors like drug use or alcohol abuse are underreported, some high-risk patients may not be flagged for additional support, which means that the program for TB treatment fails. In addition, Self-reported data may not capture other important predictors, such as biological markers or healthcare system factors (e.g., clinic accessibility or treatment side effects). While patient-reported data provides valuable context, decisions should ideally combine self-reports with objective measures (e.g., clinical records, lab results) to ensure a more complete understanding of patients' risks.

For recommendations, I would suggest that the self-reported data should be supplemented with Medical records, Clinical assessments, Social worker evaluations and Objective measures of socioeconomic status. This would essentially means we have a hybrid approach combining Self-reported demographic data (likely reliable), Professional assessment of risk factors (substance use, poverty), Clinical indicators and Social support needs assessment.While self-reported data provides valuable insights, it should not be the sole basis for determining support needs. A more comprehensive assessment approach would better identify patients requiring additional support during treatment.