Example Manuscript Template for a Data Analysis Project

Doreen Kibuule Kalembe

Warning: package 'here' was built under R version 4.4.2

**Authors**

* First Autor (ORCID: 0000-0000-1234-5678)
* Second Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.

These authors contributed equally to this work.

Corresponding author: kelvinbwambale8@gmail.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

TITLE: Sex-based analysis/Differences in TB Immune Activation.

This project is aimed at looking at the differences in the Tuberculosis immune activation between males and females.( Sex-based analysis/Differences in TB Immune Activation Research Question)

#Methods

We shall use data from a tuberculosis public repository. • Data source: Tidyverse https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/AYOFEU with a Sample size of 60 individuals. The study had 30 males and 30 females. All these participants were hiv negative and didnot have tuberculosis at the time of enrollment.

The key exposure varibles are demographic and clinicalk. These include; Key variables: - CD4 Immune activation count - Disease severity indicators(Extent of lung disease andNumber of lungs involved) -Fat in kgs - Sex/Gender - BMI - Age of the participant

#Proposed Analysis Method.

• The annalysis was conducted in R version 4.4.2. This was done to Compare immune activation levels between males and females.

We computed the male-to-female differences in disease severity using cd4 immune activation counts.We correlated immune markers with disease severity measures/Correlation between activation levels and disease severity. We did this using linear regression and machine learning models to predict which sex will show the most results. We also adjusted for potential confounders like age.

#Results

#conclusions. Sex differences in TB is contributed by more than just hormonal influence. Our study shows that TB is more severe in men than it is in women. Regardless of the medical seeking behaviors of women, women tend to a have a high body fat percentage more than men.The higher the fat percentage an individual has, the lower the risk of progression of the tuberculosis infection to disease hence lowering mortality rates among females compared to men. In addition, women tend to have more lean body mass compared to men hence having a lower chance of suffering from adverse effects of TB compared to men.

#load packages  
library("readxl")  
library("readr")  
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

library("ggplot2")  
library("tidyverse")

Warning: package 'tidyverse' was built under R version 4.4.2

Warning: package 'purrr' was built under R version 4.4.3

Warning: package 'stringr' was built under R version 4.4.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ lubridate 1.9.3 ✔ tibble 3.2.1  
✔ purrr 1.0.4 ✔ tidyr 1.3.1

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("here")  
  
# follow through with the code folder to see the #variables in the data and how i selected the #variables make this data set.   
#location of the data  
data\_location <- here::here("data","processed-data","Tb\_immune\_data.rds")  
  
#loading the data  
data<- read\_rds(data\_location)  
  
#summary of the data  
summary(data)

ID Sex Participant age BMI   
 Min. :50004 Length:61 Min. :17.00 Min. :15.55   
 1st Qu.:50020 Class :character 1st Qu.:23.00 1st Qu.:18.09   
 Median :50035 Mode :character Median :28.00 Median :19.25   
 Mean :50035 Mean :30.02 Mean :19.83   
 3rd Qu.:50050 3rd Qu.:35.00 3rd Qu.:21.47   
 Max. :50067 Max. :70.00 Max. :28.15   
 NA's :1 NA's :1   
 Dose Fat in kg LBM in kg Fat %   
 Min. :29.61 Min. :-96.640 Min. :-41.98 Min. :-177.647   
 1st Qu.:29.99 1st Qu.: 3.554 1st Qu.: 39.13 1st Qu.: 6.629   
 Median :30.00 Median : 8.704 Median : 43.43 Median : 16.685   
 Mean :29.99 Mean : 9.615 Mean : 43.02 Mean : 17.950   
 3rd Qu.:30.00 3rd Qu.: 13.364 3rd Qu.: 49.82 3rd Qu.: 25.987   
 Max. :30.15 Max. : 80.979 Max. :151.04 Max. : 207.639   
 NA's :1 NA's :1 NA's :1 NA's :1   
 LBM in % CD4+ CD4 Immune activation count  
 Min. :-107.64 Min. : 8.636 Min. : 84.0   
 1st Qu.: 74.01 1st Qu.:185.006 1st Qu.: 914.8   
 Median : 83.32 Median :317.941 Median : 1384.5   
 Mean : 82.05 Mean :337.364 Mean : 1948.1   
 3rd Qu.: 93.37 3rd Qu.:484.889 3rd Qu.: 2057.5   
 Max. : 277.65 Max. :785.896 Max. :17872.0   
 NA's :1 NA's :1 NA's :1

# 2. Introduction

## 2.1 General Background Information

Tuberculosis (TB) remains a significant global public health challenge. According to the World Health Organization (WHO), reports show that 10 million new cases and 1.4 million deaths in 2023 (1) were as a result of Tuberculosis. The spread of M. tuberculosis, as the pathogen of TB, has long been hypothesized to occur more often in the household than in the community(2). While most of the M.Tuberculosis transmission in the community is attributed to the unknown contact networks that occurs between individuals, this accounts upto 90% of the transmission. The recent outbreak of Tuberculosis in the united states clearly explains the contact network phenomena. Overall, TB accounts for one quarter of the annual deaths that happen worldwide. An estimated 10.0 million people fell ill with TB in 2018. Geographically, Southeast Asia ranked highest (44%), followed by Africa (24%) and the Western Pacific (18%), with lesser percentages in the Eastern Mediterranean (8%), the Americas (3%), and Europe (3%). Among women of reproductive years, TB is the largest cause of death (3). Global Strategies made by WHO, aim at ending the TB epidemic, with targets to reduce the number of deaths caused by TB by 95% and to decrease new cases by 90% from 2015 to 2035(4).

Mycobacterium tuberculosis infection primarily spreads when a person shares airspace with someone who has active tuberculosis and inhales infectious droplet aerosols. Since the bacterium transmits through respiratory routes to susceptible individuals, the household of an infected person becomes a critical environment for intense M. tuberculosis transmission, where most new infections and cases can be detected and treated. Research reports indicate significant differences in tuberculosis responses between men and women. Notably, these studies suggest that women face greater barriers to early detection and treatment of tuberculosis compared to men, potentially impacting disease outcomes and progression.

The World Health Organisations, among others, attributed to the differences in the tuberculosis trends between men and women to access to care and treatment, pregnancy effects,body fat mass and gender rules that diminish the social capital of women hence lowering incidence among them(5). Not forgetting that biological mechanisms might influence these differences, previous studies show that men have a higher incidence and prevalence of disease and this might lead to bad treatment outcomes as compared to women.

Body fat plays an important role in supporting immune function, and women typically maintain higher fat storage than men. This increased fat storage may contribute to women’s stronger immune responses. Since the balance of fat and lean tissue differs significantly between men and women, particularly at the time of disease presentation, the objective of this study is to determine whether higher immune activation correlates with greater infection severity or illness.

## 2.2 Description of data and data source

The data we used for this project was got from the TRAC dataverse and is fully accessible to everyone. This is considered a safe place where the the Tuberculosis data from the TRAc project is kept and can be used by anyone who wants contribute to TB research. More details about the data can be found [here]!(https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/AYOFEU). The data is publicaly available and can be accessed by anyone.

## 2.3 Questions/Hypotheses to be addressed

Since the balance of fat and lean tissue differ in men and women especially at the time of tuberculosis presentation, the objective of this study will be to find out whether there are differences in the Tuberculosis immune activation between males and females. We shall hypothesize on the more immune activation there is, the more infected / sicker the person.

# 3. Methods

Study Design

This study utilized a cross sectional study design analyzing data from 2017.Sixty TB-confirmed participants (30 males, 30 females, aged 15+) were recruited from Mulago Hospital.

Our study included patients aged 15 years and older who were experiencing their first episode of tuberculosis. To maintain data integrity, we excluded individuals who were receiving hormonal therapy or using hormonal contraceptives. Additionally, patients with comorbidities such as diabetes mellitus, cancer, or HIV were excluded from the study. Conversely, we applied strict exclusion criteria to ensure a focused research population: all individuals under 15 years of age, those with recurrent TB episodes, patients taking any form of hormonal therapy or contraception, and those diagnosed with diabetes mellitus, cancer, or HIV were not enrolled in the study.

The variables of interest were extracted from the TRAC dataset and exported to an MS Excel sheet. The outcome variable was CD4+ immune activatin counts. The primary independent variable was sex (cis-gender) and this was categorized as male and female. The covariates were age at diagnosis,BMI, Fat in Kg, lean body mass in kg(LBM in kg), CD4+ and CD4 immune activation counts. other variables like fat mass index (FMI) and fat-free mass index (FFMI) were also obtained.

## 3.1 Schematic of workflow

Chromosomal sex and hormones play a role in the modulation of immune responses to Mtb and thereby contribute to the general outcome of disease. The first cells that encounter Mtb are known as Macrophages. These initiate a local inflammatory reaction which may differ between the sexes(6). This causes the long term containment of TB infection to differ between men and women since their adaptive immune responses are different. (**tbfig-1?**)

|  |
| --- |
| Figure 1: Impact of sex differences on the immune response to Mtb and their influence disease outcome. |

## 3.2 Data import and cleaning

We imported the data from MS Excel to R. We explored the data and found some missing values for some variables. We decided to drop one observation since the person lacked the primary exposure and did not have most of the details we needed for our final analysis. We categorized the participants age into four categories. That is 17-<28, 28-<40, 40-<60 and >=60.

## 3.3 Statistical analysis

The final cleaned data set contains some negative fat values. Negative values occur when the deuterium dose has not had sufficient time to fully equilibrate with body water, or the dose was not completely consumed. This is seen as low deuterium enrichment, resulting in an overestimation of the size of the body water pool and hence high FFM (FFMI = LBM / height2)and low % body fat.

We applied three supervised machine learning models — linear regression, LASSO regression, and Random Forests — to predict CD4 Immune Activation Count based on Sex, Age, BMI, Lean Body Mass (LBM), Fat mass, and CD4+ cell count. Data preprocessing included median imputation for missing values, dummy encoding of categorical variables, normalization, and removal of highly correlated features. Models were evaluated using RMSE and visualized through observed vs. predicted plots. Hyperparameter tuning was performed using grid search with cross-validation.

The significance of Sex as a predictor of immune activation was assessed in each model. For linear models, coefficients for Sex were examined along with their corresponding p-values. The Random Forest model’s variable importance was assessed to identify the predictive contribution of Sex.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

We explored the data and using figures and summary tables. We used box plots, histograms and and correlation plots to see whether we can find something interesting in the data. Scatter plots were used to see the trends in the data and also see whether there are some interesting relationships between the continuous variables in the data.

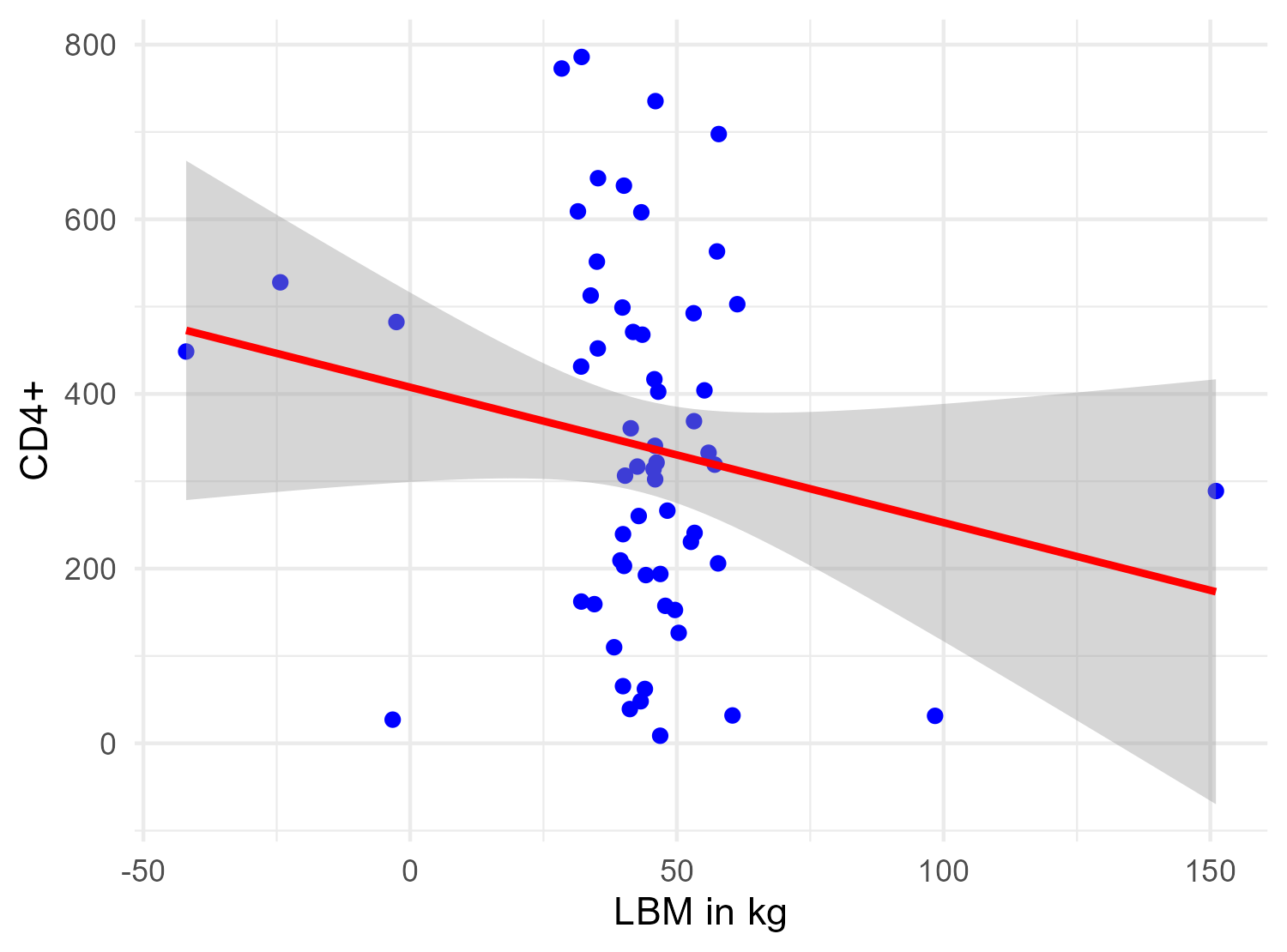
[Table 1](#tbl-TABLE1) shows a summary of the data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data descriptive summary table.   |  | F | M | Overall | | --- | --- | --- | --- | |  | (N=30) | (N=30) | (N=60) | | CD4+ |  |  |  | | Mean (SD) | 394 (217) | 289 (181) | 341 (205) | | Median [Min, Max] | 449 [27.0, 786] | 278 [8.64, 735] | 319 [8.64, 786] | | Missing | 1 (3.3%) | 0 (0%) | 1 (1.7%) | | CD4 Immune activation count |  |  |  | | Mean (SD) | 1750 (1340) | 2160 (3170) | 1960 (2430) | | Median [Min, Max] | 1520 [94.0, 7350] | 1210 [84.0, 17900] | 1380 [84.0, 17900] | | Missing | 1 (3.3%) | 0 (0%) | 1 (1.7%) | | Participant age |  |  |  | | Mean (SD) | 29.9 (12.8) | 30.2 (6.83) | 30.0 (10.1) | | Median [Min, Max] | 26.0 [17.0, 70.0] | 28.5 [19.0, 48.0] | 28.0 [17.0, 70.0] | | Fat in kg |  |  |  | | Mean (SD) | 18.9 (22.4) | 0.348 (21.9) | 9.62 (23.8) | | Median [Min, Max] | 10.9 [-0.192, 81.0] | 4.37 [-96.6, 21.3] | 8.70 [-96.6, 81.0] | | LBM in kg |  |  |  | | Mean (SD) | 31.1 (21.0) | 54.9 (21.1) | 43.0 (24.1) | | Median [Min, Max] | 39.6 [-42.0, 45.9] | 50.0 [38.2, 151] | 43.4 [-42.0, 151] | |

## 4.2 Basic statistical analysis

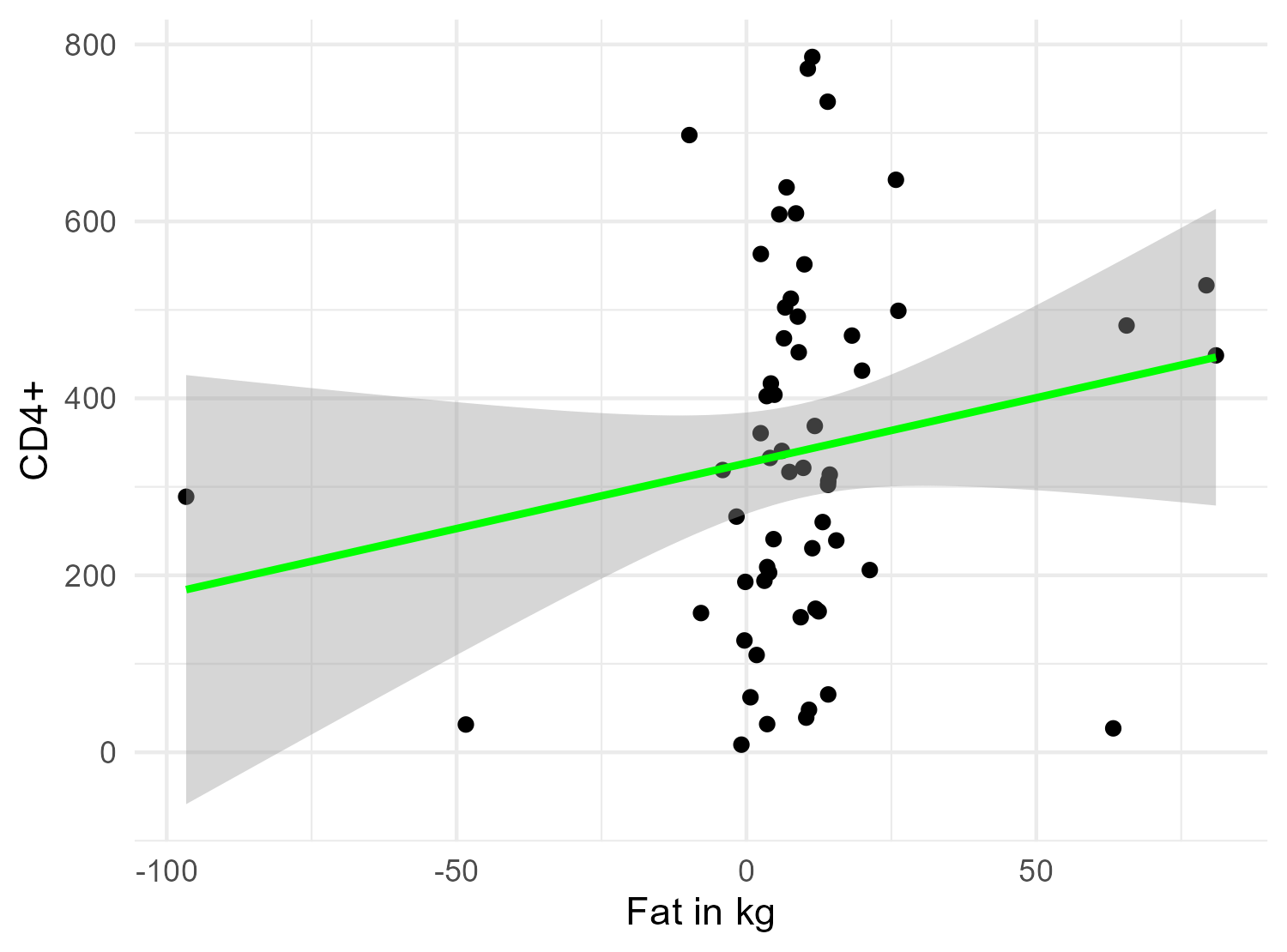
We used simple linear model to inspect whether there was an association between our main outcome CD4 immune activation count and the the primary predictor sex. To improve the model, we added in other predictors that could influence the immune activation count by performing a multi-variant linear regression anlaysis.

(**fig1-result?**) shows a scatter plot of lean body mass in kg and CD4



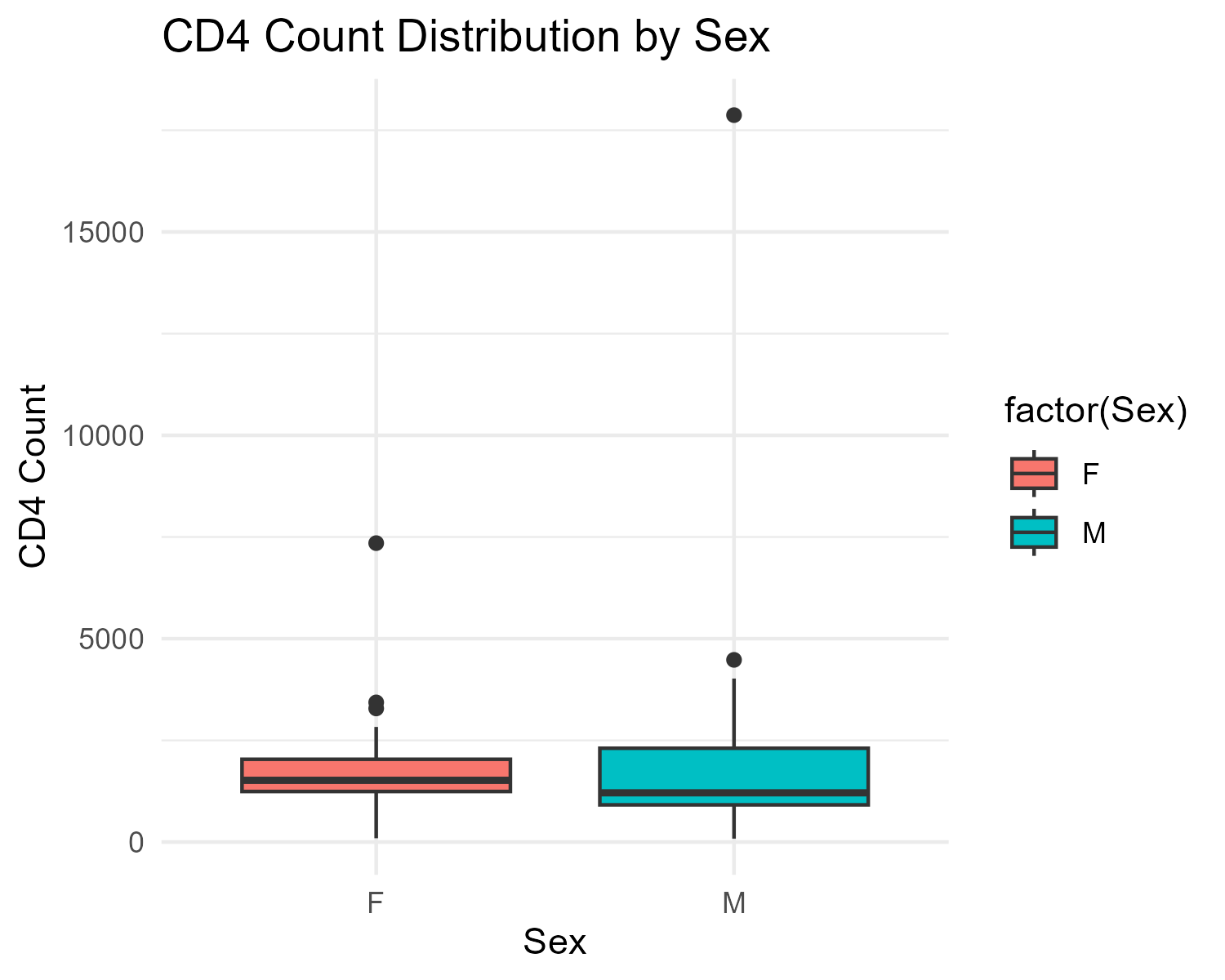
Scatter plot of CD4+ cells versus LBM in kg.

(**fig2-result?**) shows a scatter plot of CD4 vs Fat in kg.



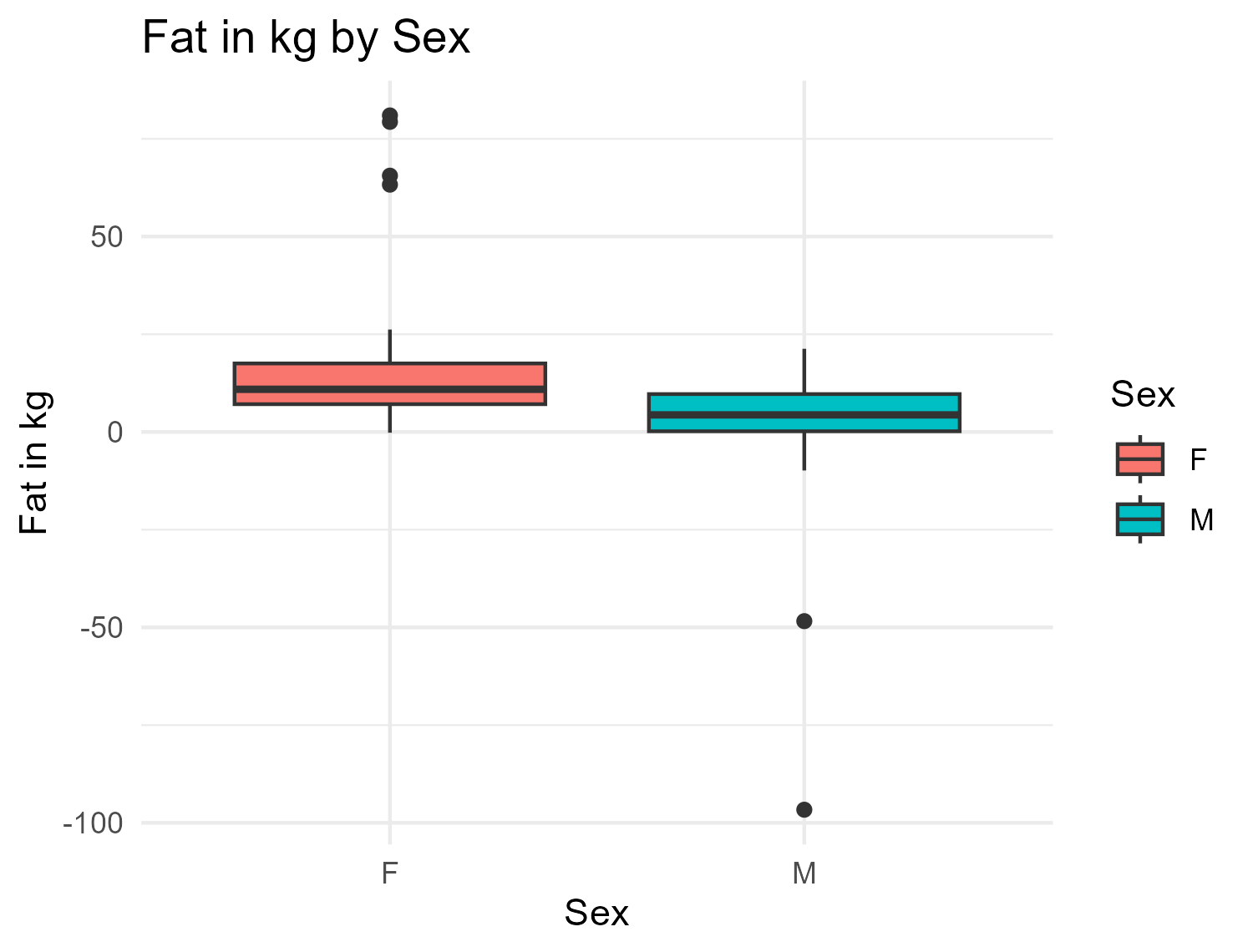
Scatter plot of CD4+ cells versus Fat in kg.

(**plot1-result?**) a boxplot of the cd4 immune activation count by sex.



Box plot to show the distribution of cd4 count immune activation counts among males and females.

(**plot2-result?**) a boxplot to show the distributions of fat in kilograms among males females.



Box plot showing Fat in kg among males and females.

## 4.3 Full analysis

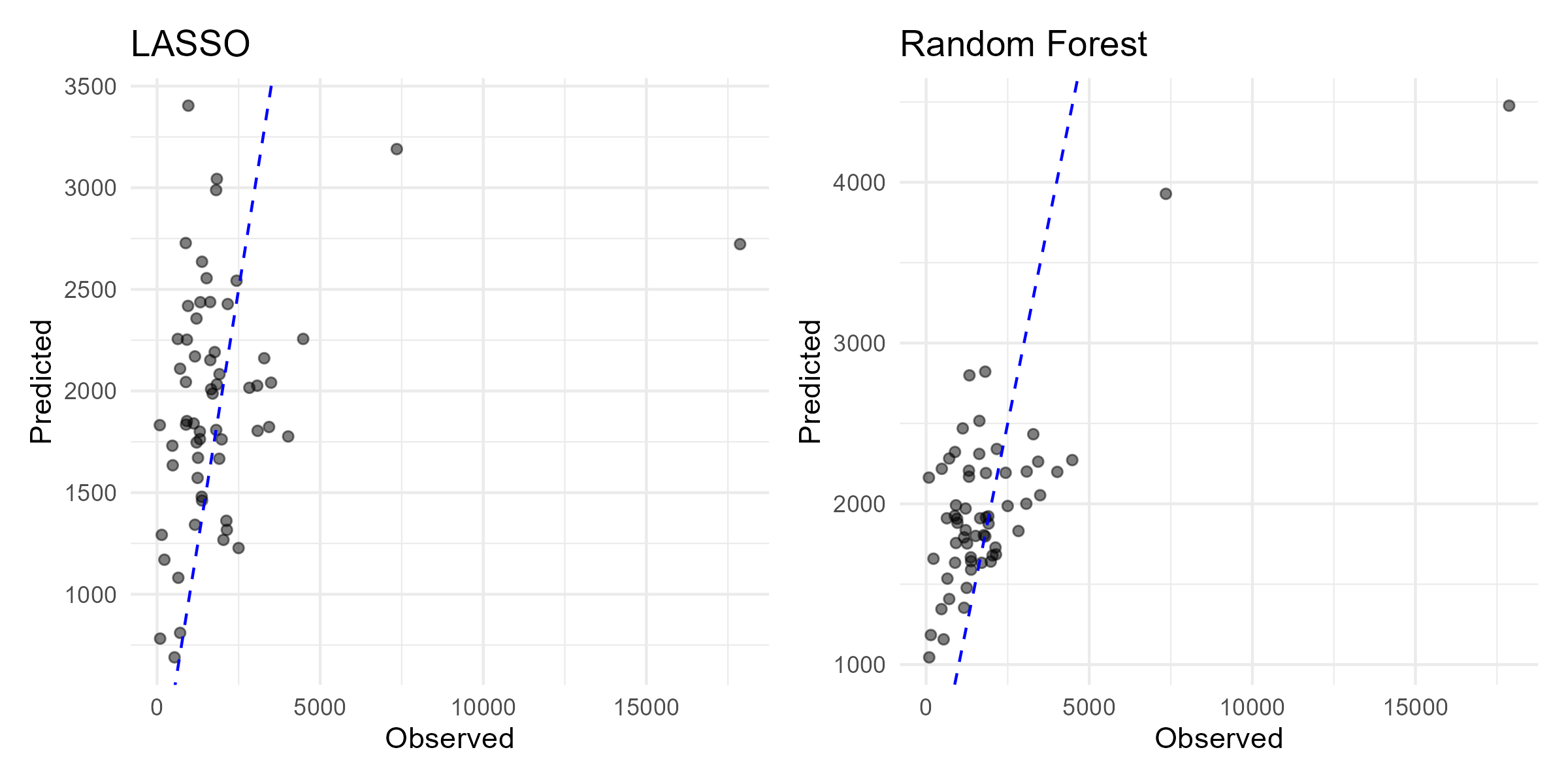
In order to take care of confounding, we performed a multi-regression analysis and the results of the analysis are summarized in the table below.

This [Table 2](#tbl-model_table2) shows a summary of a multi-variant linear regression model fit.

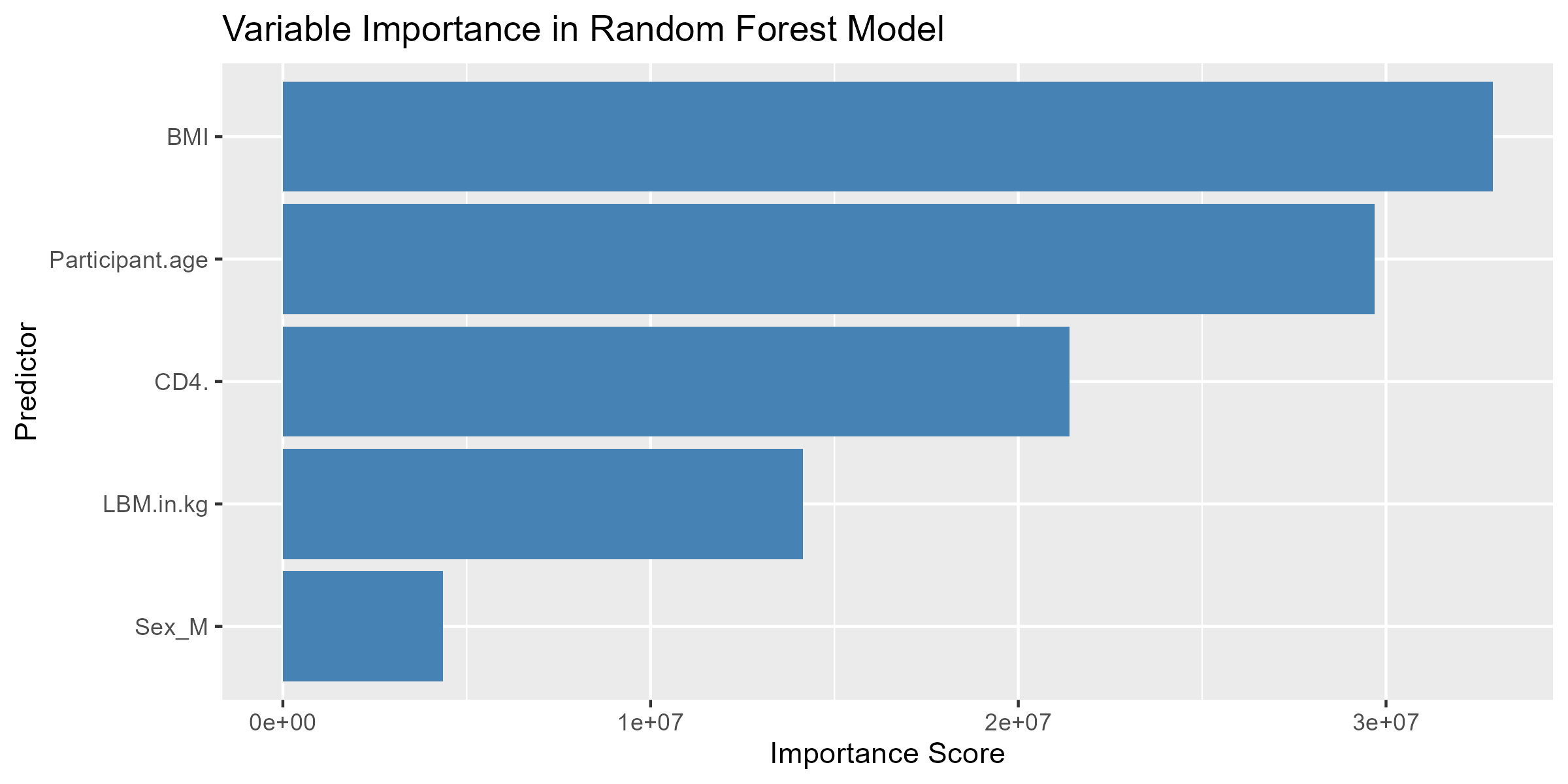
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model summary table 2.   |  | Estimate | Std. Error | t value | Pr(>|t|) | | --- | --- | --- | --- | --- | | (Intercept) | 3759.632173 | 2567.133417 | 1.4645254 | 0.1490704 | | relevel(factor(Sex), ref = “F”)M | 928.102409 | 804.136468 | 1.1541603 | 0.2537114 | | Participant age | -6.818317 | 31.975088 | -0.2132384 | 0.8319748 | | BMI | -174.848153 | 177.081620 | -0.9873874 | 0.3280250 | | LBM in kg | 3.933569 | 60.703367 | 0.0647998 | 0.9485818 | | Fat in kg | 13.122848 | 58.955129 | 0.2225904 | 0.8247263 | | CD4+ | 3.215944 | 1.623752 | 1.9805632 | 0.0529418 | |

We finally considered using ML algorithm to fit our data. The Random Forest model achieved the lowest RMSE (1132), outperforming both the linear regression and LASSO models, which had RMSEs around 2107. This figure (**Predicted\_vs\_observed?**) shows a comparison of observed vs. predicted values for each model.

This figure (**Variable\_plot?**) shows the variable importance plot from the random forest model. This highlights the most influential predictors of CD4 immune activation. BMI, participants age, CD4+, Lean body mass(LBM) and Sex emerged as the most important features, indicating they contributed most to the model’s predictive accuracy. This suggests that body composition and biological sex may play a key role in immune activation patterns.



predicted vs observed of three models



Variable plot from the random forests model.

We applied LASSO regression to identify preredictors of CD4 immune activation. After removing the participant ID variable, the model retained BMI (estimate = -297), CD4+ count (estimate = 514), sex (male) (estimate = 260), and lean body mass (estimate = -22.7) as non-zero coefficients. Participant age was excluded from the final model.

Interestingly, LBM (Lean Body Mass) and Participant Age were ranked lower, suggesting they had less influence in this dataset. These findings align with the Random forest results, which also retained BMI,CD4+ lean body mass and sex as key predictors.

# 5. Discussion

In this analysis of sex-based differences in TB-related immune activation, we aimed to identify key predictors of CD4 immune activation count , which we used as a proxy for severity of disease. Our hypothesis was that higher immune activation counts indicate more severe illness, and that biological sex may influence immune responses to TB.

when we used both the LASSO regression and random forest models, we identified CD4+ count, BMI, sex, and lean body mass (LBM) as important predictors of immune activation. Our findings highlight a notable difference in CD4 immune activation count rates between males and females, with males exhibiting higher risk of TB infection and disease than females. Sex was retained as a significant predictor, with a positive coefficient (estimate = 260). This indicated that males experienced a 260 increase in CD4 immune activation count compared to females, who served as the reference group. This suggests that males may experience more severe immune activation, supporting the hypothesis that sex-based biological differences contribute to disease severity in TB. This result aligns with previous studies that have reported sex-based disparities in health outcomes and mortality risks as a result of TB(6),(7). Some studies still attribute the increased risk among males to be as a result of biological differences or differences in the accessibility of resources, especially healthcare. This is generally because women have better prognosis than men and they have better health seeking behavior as compared to their male counterparts.

Research shows that most women have higher fat repositories than men(8) and some studies show that the higher the fat percentage an individual has, the lower the risk of progression of the tuberculosis infection to disease hence lowering mortality rates among females compared to men. In addition, women tend to have more lean body mass compared to men hence having a lower chance of suffering from adverse effects of TB compared to men.

## 5.1 Strengths and Limitations

This study has several strengths. We applied simple linear regression, LASSO regression and random forest models for annalysis of our data. We were able to capture a range of relationships between predictors and immune activation, which in turn increased the robustness of our findings. The focus on sex-based differences in TB immune response addressed an important and often neglected question in infectious disease research especially among TB patients who are HIV negative. Additionally, the use of biologically meaningful predictors such as BMI, lean body mass, CD4+ count, and sex enhances the interpretability and relevance of the results. The consistency of key predictors across the LASSO model and the random forests model further supports the reliability of our conclusions. However, there are limitations to consider. The relatively small sample size may reduce the generalizability and stability of the findings. As an observational study, we are limited to identifying associations rather than making causal claims. Important clinical variables, such as TB stage or treatment history, may not have been included, potentially introducing residual confounding. Moreover, while CD4 immune activation was used as our marker for disease severity, it may not fully capture the clinical picture.

Despite limitations of the present of our study, some studies show that Patients with wasting do not recover to the same levels of body mass as those who were not wasted. This suggests that tuberculosis may leads to permanent loss of lean tissue and fat mass.

## 5.2 Conclusions

Our results support the hypothesis that higher immune activation is associated with greater disease severity, and demonstrate that males exhibit higher CD4 immune activation levels compared to females, indicating they may experience more intense immune responses and possibly more severe disease outcomes. These findings highlight the importance of considering sex-based biological differences in TB research and clinical management.

In addition to sex, BMI and lean body mass emerged as important predictors, suggesting that body composition may influence the immune landscape in TB. These insights highlight the need for sex-specific approaches in TB monitoring and treatment, and point to CD4 immune activation as a potentially valuable biomarker for assessing disease severity. However, the need for future research is needed to validate these findings especially among larger populations to explore the underlying sex mechanism and body composition differences in regard to immune activation.

# 6. References

1. Santos M de LSG, Vendramini SHF, Gazetta CE, et al. Poverty: Socioeconomic characterization at tuberculosis. *Revista Latino-Americana de Enfermagem* [electronic article]. 2007;15(spe):762–767. (<http://dx.doi.org/10.1590/S0104-11692007000700008>)

2. Martinez L, Shen Y, Mupere E, et al. Transmission of mycobacterium tuberculosis in households and the community: A systematic review and meta-analysis. *American Journal of Epidemiology* [electronic article]. 2017;185(12):1327–1339. (<http://dx.doi.org/10.1093/aje/kwx025>)

3. Dye C, Garnett GP, Sleeman K, et al. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *The Lancet* [electronic article]. 1998;352(9144):1886–1891. (<http://dx.doi.org/10.1016/S0140-6736(98)03199-7>)

4. Murgia Y, Sepulcri C, Crupi L, et al. Tuberculosis infection control: Experiences and considerations from a web based tool implementation. In: *pHealth 2024*. IOS Press; 2024(<http://dx.doi.org/10.3233/SHTI240053>)

5. Whalen C. Sex differences in the presentation of tuberculosis in uganda. 2023;(<https://dataverse.unc.edu/citation?persistentId=doi:10.15139/S3/AYOFEU>)

6. Hertz D, Schneider B. Sex differences in tuberculosis. *Seminars in Immunopathology* [electronic article]. 2018;41(2):225–237. (<http://dx.doi.org/10.1007/s00281-018-0725-6>)

7. Marçôa R, Ribeiro AI, Zão I, et al. Tuberculosis and gender – factors influencing the risk of tuberculosis among men and women by age group. *Pulmonology* [electronic article]. 2018;24(3):199–202. (<http://dx.doi.org/10.1016/j.pulmoe.2018.03.004>)

8. Karastergiou K, Smith SR, Greenberg AS, et al. Sex differences in human adipose tissues – the biology of pear shape. *Biology of Sex Differences* [electronic article]. 2012;3(1). (<http://dx.doi.org/10.1186/2042-6410-3-13>)