FACTORS ASSOCIATED WITH GENOMIC ALTERATIONS IN TUMOR SAMPLES

Group 10 Capstone Project Report (Modeling Part)

Yuxiang Ren, Amanda Carrico

2024-09-24

Introduction

Genomic alterations in tumors have been found to be prognostic and are increasingly being used in treatment decision making. One way of measuring this is tumor mutation burden (TMB). Tumor mutation is the total number of mutations found in the DNA (mutations per megabase) of the cancer cells within a sample. Tumor tissue is analyzed for this via various methods depending on the sequencing technology being used. There were three methods relevant to our research. Whole exome sequencing will measure all of the protein coding regions of the genome (the exome) so the mutations being counted would only be within this region. In this case, TMB is calculated by counting the number of mutations and dividing by the total number of megabases in the exome. Whole exome sequencing allows for comprehensive analysis of the parts of the genome that will have impact on gene expression while allowing for a good cost balance, though there is the limitation of solely exoms. Targeted gene panels focus on specific genes such as those already known to be associated with cancer. Thus, TMB is the the number of mutations within this subset divided by the total number of megabases sequenced (this subset). The method is focused on genes that will be involved which makes it cheaper but could also lead to bias as it is not very comprehensive. Whole genome sequencing is when the whole genome is sequenced to have TMB from the total number of mutations divided by the number of megabases in the entire genome. This is very comprehensive but high in cost. Each method has its strengths and weaknesses. For our data, most of the studies we investigated used whole exome sequencing.

According to the “Mutation Burden Independently Predicts Survival in the Pan-Cancer Atlas” study, studies on past tumor mutation do not show a linear relationship between TMB and survival. It is instead seen, after fitting a quadratic model, that “patients with intermediate TMB levels had a significantly poorer survival prognosis than patients with either low or high TMB” (“Mutation Burden Independently Predicts Survival in the Pan-Cancer Atlas”). This study found, for multiple cancer types, that after a certain threshold, TMB is associatied with reduced mortal hazard. These effects dependent on TMB are also seen in clinical efficacy of treatments such as immune checkpoint inhibitors. In the “Tumor mutation burden predicts response and survival to immune checkpoint inhibitors: a meta-analysis" study, High

TMB was significantly associated with better progression free survival than low tumor mutation burden patients. This finding was generalized to many cancer types but discovered that there was not one universal TMB cutoff for all cancer types in another study. The effects of TMB are clear but need to be studied more. Another genomic alteration measurement is fraction genome altered, which we will also be investigating.

Fraction genome altered (FGA) is the proportion of the genome that has had somatic alterations. This includes any form of alteration such as mutation, copy number variation, or translocation. It is calculated by first determining the regions of the genome that have been altered and the total size of the genome in base pairs. The number of base pairs involved in the altered region is divided by the total for the FGA. It, again, depends on how much of the genome is sequenced – such as the previously discussed methods of whole genome or targeted. This measurement shows a lot about the tumor and can also provide insight into treatment and prognosis of the cancer. The fraction of genomic alteration was found to be significant in predicting progression free survival and disease specific survival in "Genomic alterations predictive of poor clinical outcomes in pan-cancer". This study also suggests evidence of impact on treatment that needs to be further explored.

Tumor mutation burden and fraction genome altered have been shown to have significant effects on survival and efficacy of treatments. Due to this, it is imperative that measurements of genomic alteration and what impacts them are studied. There are many factors that could potentially affect the values of these genomic alteration measures. Our research aims to investigate this. We are specifically interested in whether sex, age, and smoking history status have an association with tumor mutation burden and fraction genome altered for various cancer types. We also hope to assess any impact on genomic alterations based treatment decision making and outcome prediction. We will conduct individual factor analysis and descriptive statistics as well as group factor analysis to study the relationships between all variables and the genomic alteration measurement. This and further study in the area could lead to better treatment decision making and understanding in outcome prediction.

Methods

Missing Value Imputation

```
## Load combined database
Full_smoking <-
read.csv("/Users/JasonRen.584/Desktop/capstone/amd/08:28/Full_smoking.csv")
# data  =  datasets input for modeling
data <- Full_smoking
head(data)
```

```
##         Study PATIENT_ID     SAMPLE_ID        CANCER_TYPE
SITE_OF_TUMOR_TISSUE
## 1 msk_ch_2020  P-0000004 P-0000004-N01      Breast Cancer
<NA>
## 2 msk_ch_2020  P-0000015 P-0000015-N01      Breast Cancer
<NA>
## 3 msk_ch_2020  P-0000023 P-0000023-N01       Mesothelioma
<NA>
## 4 msk_ch_2020  P-0000024 P-0000024-N01 Endometrial Cancer
<NA>
## 5 msk_ch_2020  P-0000025 P-0000025-N01 Endometrial Cancer
<NA>
## 6 msk_ch_2020  P-0000026 P-0000026-N01 Endometrial Cancer
<NA>
##      SEX  RACE      AGE OS_STATUS OS_MONTHS SMOKING_HISTORY TUMOR_PURITY
## 1 Female White 39.73990        NA        NA            Never           NA
## 2 Female White 44.44079        NA        NA            Never           NA
## 3   Male White 61.31964        NA        NA            Never           NA
## 4 Female White 61.34428        NA        NA           Former           NA
## 5 Female White 72.67351        NA        NA           Former           NA
## 6 Female Asian 71.70979        NA        NA           Former           NA
##   METASTATIC        TMB FGA
## 1       <NA> 0.06666667  NA
## 2       <NA>         NA  NA
## 3       <NA>         NA  NA
## 4       <NA>         NA  NA
## 5       <NA>         NA  NA
## 6       <NA>         NA  NA
```

```r
# check missing values
missing_values <- colSums(is.na(data))
print(missing_values)
```

```
##                Study           PATIENT_ID            SAMPLE_ID
##                    0                    0                    0
##          CANCER_TYPE SITE_OF_TUMOR_TISSUE                  SEX
##                  102                24321                    0
##                 RACE                  AGE            OS_STATUS
##                    0                 5825                23663
##           OS_MONTHS      SMOKING_HISTORY         TUMOR_PURITY
##               24983                    0                25671
##           METASTATIC                  TMB                  FGA
##                26919                16619                23233
```

In the statistical modeling phase, the primary data set used is named “Full Smoking,” which was created during the data processing stage.Due to its size and optimization of non missing data among all categories, the smoking data set was used in further analyses with various models. This dataset includes the following variables: Study, Patient_ID, Sample_ID, cancer type, site of timuuor tissue, sex, race, age, os_status, os_month, smoking history, tumor purity, metastatic, Tumor Mutation Burden (TMB), and Fraction Genome Altered

(FGA). A snippet of this database is shown above. The dataset ensures that key variables such as patient_id, Study, sample_id, smoking history, sex, and race have no missing values, making it suitable for robust statistical analysis. This complete dataset forms the basis for further statistical exploration, including the construction of mixed-effects models.

```
# # pre-check the data types of all the variables
str(data)
```

```
## 'data.frame':    31602 obs. of  15 variables:
##  $ Study               : chr  "msk_ch_2020" "msk_ch_2020" "msk_ch_2020"
"msk_ch_2020" ...
##  $ PATIENT_ID          : chr  "P-0000004" "P-0000015" "P-0000023" "P-
0000024" ...
##  $ SAMPLE_ID           : chr  "P-0000004-N01" "P-0000015-N01" "P-0000023-
N01" "P-0000024-N01" ...
##  $ CANCER_TYPE         : chr  "Breast Cancer" "Breast Cancer"
"Mesothelioma" "Endometrial Cancer" ...
##  $ SITE_OF_TUMOR_TISSUE: chr  NA NA NA NA ...
##  $ SEX                 : chr  "Female" "Female" "Male" "Female" ...
##  $ RACE                : chr  "White" "White" "White" "White" ...
##  $ AGE                 : num  39.7 44.4 61.3 61.3 72.7 ...
##  $ OS_STATUS           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ OS_MONTHS           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SMOKING_HISTORY     : chr  "Never" "Never" "Never" "Former" ...
##  $ TUMOR_PURITY        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ METASTATIC          : chr  NA NA NA NA ...
##  $ TMB                 : num  0.0667 NA NA NA NA ...
##  $ FGA                 : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
### (Imputation for Missing Values)
## Impute missing categorical data  ->  unknown

# we don't impute response variables
data <- data %>%
  filter(!is.na(TMB)) %>%
  filter(!is.na(FGA))

categorical_na <-
c("OS_STATUS","CANCER_TYPE","METASTATIC","SITE_OF_TUMOR_TISSUE")

# transform missing categorical data
data <- data %>%
  mutate(across(all_of(categorical_na), ~ replace_na(as.character(.),
"unknown"))) %>%
  mutate(across(all_of(categorical_na), as.factor))
# head(data)
```

Since our final dataset is derived from eight different selected studies, the variables measured in each study are not identical. As a result, our merged database contains a substantial amount of missing values. Therefore, it is necessary to perform appropriate

imputation of these missing values before modeling. First, we excluded records with missing values for the response variables (TMB and FGA). Subsequently, we assigned a unified level, 'unknown,' to both missing values and invalid values across all categorical variables.

```
# # impute missing numeric data

imputed_data <- mice(data, m = 5, method = 'pmm', maxit = 50, seed = 500)

##
##  iter imp variable
##   1   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   1   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   1   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   1   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   1   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   2   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   2   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   2   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   2   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   2   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   3   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   3   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   3   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   3   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   3   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   4   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   4   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   4   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   4   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   4   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   5   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   5   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   5   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   5   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   5   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   6   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   6   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   6   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   6   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   6   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   7   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   7   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   7   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   7   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   7   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   8   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   8   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   8   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   8   4  AGE*  OS_MONTHS*  TUMOR_PURITY
```

```
##    8   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##    9   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##    9   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##    9   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##    9   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##    9   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   10   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   10   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   10   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   10   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   10   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   11   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   11   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   11   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   11   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   11   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   12   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   12   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   12   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   12   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   12   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   13   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   13   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   13   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   13   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   13   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   14   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   14   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   14   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   14   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   14   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   15   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   15   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   15   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   15   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   15   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   16   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   16   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   16   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   16   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   16   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   17   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   17   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   17   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   17   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   17   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   18   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   18   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   18   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   18   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
```

```
##   18   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   19   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   19   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   19   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   19   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   19   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   20   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   20   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   20   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   20   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   20   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   21   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   21   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   21   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   21   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   21   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   22   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   22   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   22   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   22   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   22   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   23   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   23   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   23   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   23   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   23   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   24   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   24   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   24   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   24   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   24   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   25   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   25   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   25   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   25   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   25   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   26   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   26   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   26   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   26   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   26   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   27   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   27   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   27   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   27   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   27   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   28   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   28   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   28   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   28   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
```

```
##   28   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   29   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   29   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   29   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   29   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   29   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   30   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   30   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   30   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   30   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   30   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   31   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   31   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   31   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   31   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   31   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   32   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   32   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   32   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   32   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   32   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   33   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   33   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   33   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   33   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   33   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   34   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   34   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   34   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   34   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   34   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   35   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   35   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   35   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   35   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   35   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   36   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   36   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   36   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   36   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   36   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   37   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   37   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   37   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   37   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   37   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   38   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   38   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   38   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   38   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
```

```
##   38   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   39   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   39   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   39   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   39   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   39   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   40   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   40   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   40   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   40   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   40   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   41   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   41   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   41   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   41   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   41   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   42   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   42   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   42   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   42   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   42   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   43   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   43   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   43   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   43   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   43   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   44   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   44   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   44   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   44   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   44   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   45   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   45   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   45   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   45   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   45   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   46   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   46   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   46   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   46   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   46   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   47   1  AGE*  OS_MONTHS*  TUMOR_PURITY
##   47   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   47   3  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   47   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   47   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   48   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   48   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   48   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   48   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
```

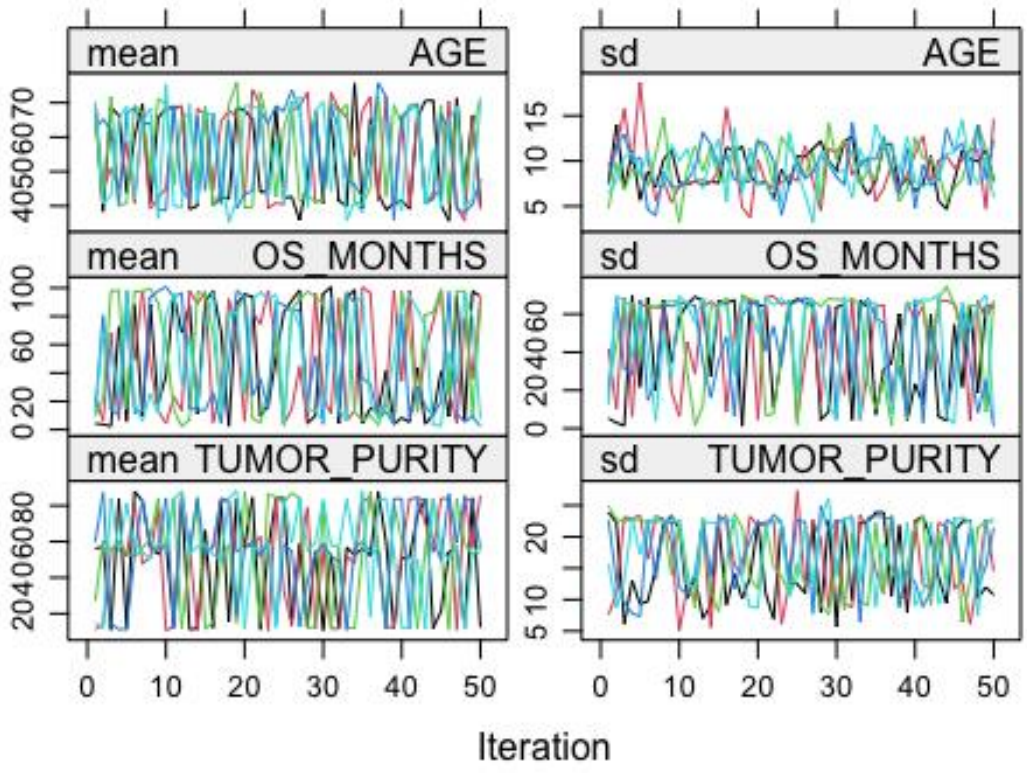
```
##   48   5  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   49   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   49   2  AGE*  OS_MONTHS*  TUMOR_PURITY
##   49   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   49   4  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   49   5  AGE*  OS_MONTHS*  TUMOR_PURITY
##   50   1  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   50   2  AGE*  OS_MONTHS*  TUMOR_PURITY*
##   50   3  AGE*  OS_MONTHS*  TUMOR_PURITY
##   50   4  AGE*  OS_MONTHS*  TUMOR_PURITY
##   50   5  AGE*  OS_MONTHS*  TUMOR_PURITY

## Warning: Number of logged events: 1396
```

```
summary(imputed_data)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##                Study           PATIENT_ID            SAMPLE_ID
##                   ""                   ""                   ""
##          CANCER_TYPE SITE_OF_TUMOR_TISSUE                  SEX
##                   ""                   ""                   ""
##                 RACE                  AGE            OS_STATUS
##                   ""                "pmm"                   ""
##            OS_MONTHS      SMOKING_HISTORY         TUMOR_PURITY
##                "pmm"                   ""                "pmm"
##           METASTATIC                  TMB                  FGA
##                   ""                   ""                   ""
## PredictorMatrix:
##                      Study PATIENT_ID SAMPLE_ID CANCER_TYPE
## Study                    0          0         0           1
## PATIENT_ID               0          0         0           1
## SAMPLE_ID                0          0         0           1
## CANCER_TYPE              0          0         0           0
## SITE_OF_TUMOR_TISSUE     0          0         0           1
## SEX                      0          0         0           1
##                      SITE_OF_TUMOR_TISSUE SEX RACE AGE OS_STATUS OS_MONTHS
## Study                                   1   0    0   1         1         1
## PATIENT_ID                              1   0    0   1         1         1
## SAMPLE_ID                               1   0    0   1         1         1
## CANCER_TYPE                             1   0    0   1         1         1
## SITE_OF_TUMOR_TISSUE                    0   0    0   1         1         1
## SEX                                     1   0    0   1         1         1
##                      SMOKING_HISTORY TUMOR_PURITY METASTATIC TMB FGA
## Study                              0            1          1   1   1
## PATIENT_ID                         0            1          1   1   1
## SAMPLE_ID                          0            1          1   1   1
## CANCER_TYPE                        0            1          1   1   1
## SITE_OF_TUMOR_TISSUE               0            1          1   1   1
```

```
## SEX                                  0            1          1   1   1
## Number of logged events:  1396
##   it im dep     meth             out
## 1  0  0     constant           Study
## 2  0  0     constant      PATIENT_ID
## 3  0  0     constant       SAMPLE_ID
## 4  0  0     constant             SEX
## 5  0  0     constant            RACE
## 6  0  0     constant SMOKING_HISTORY
```

```
plot(imputed_data)
```




```
# Data sets for manual extraction of random effects
completed_data_1 <- complete(imputed_data, 1)
completed_data_2 <- complete(imputed_data, 2)
completed_data_3 <- complete(imputed_data, 3)
completed_data_4 <- complete(imputed_data, 4)
completed_data_5 <- complete(imputed_data, 5)
# head(completed_data_1)
```

For the most critical missing numeric values, we applied multiple imputation using the Predictive Mean Matching (PMM) method via the mice package. We generated five imputed datasets with a maximum of 50 iterations. This approach ensured that the imputed data

retained as much of the original characteristics as possible while maintaining a sufficiently large data set size. The subsequent statistical modeling analysis will integrate the results from these five imputed datasets to ensure the robustness and reliability of the findings. After 50 iterations, the means and standard deviations of the variables stabilized, indicating that the imputation results had converged.

```
# post- check missing value
# factorize the categorical variables (character -> factor)
# These database is for
cleaned_data_1 <- completed_data_1 %>%
  mutate(across(where(is.character), as.factor))
cleaned_data_2 <- completed_data_2 %>%
  mutate(across(where(is.character), as.factor))
cleaned_data_3 <- completed_data_3 %>%
  mutate(across(where(is.character), as.factor))
cleaned_data_4 <- completed_data_4 %>%
  mutate(across(where(is.character), as.factor))
cleaned_data_5 <- completed_data_5 %>%
  mutate(across(where(is.character), as.factor))
cleaned_data <- completed_data_1 %>%
  mutate(across(where(is.character), as.factor))

head(cleaned_data)
```

```
##             Study PATIENT_ID         SAMPLE_ID        CANCER_TYPE
## 1 msk_impact_2017  P-0000015 P-0000015-T01-IM3      Breast Cancer
## 2 msk_impact_2017  P-0000023 P-0000023-T01-IM3       Mesothelioma
## 3 msk_impact_2017  P-0000025 P-0000025-T01-IM3 Endometrial Cancer
## 4 msk_impact_2017  P-0000025 P-0000025-T02-IM5 Endometrial Cancer
## 5 msk_impact_2017  P-0000026 P-0000026-T01-IM3 Endometrial Cancer
## 6 msk_impact_2017  P-0000027 P-0000027-T01-IM3       Mesothelioma
##   SITE_OF_TUMOR_TISSUE    SEX    RACE AGE OS_STATUS OS_MONTHS
SMOKING_HISTORY
## 1               Breast Female Unknown  46         1  71.10454
Never
## 2           Peritoneum   Male Unknown  47         1   8.71000
Never
## 3               Uterus Female Unknown  47         0   8.81000
Never
## 4               Uterus Female Unknown  33         0   8.81000
Never
## 5               Uterus Female Unknown  33         0  71.10454
Never
## 6                 Lung Female Unknown  47         1 203.35306
Never
##   TUMOR_PURITY METASTATIC      TMB    FGA
## 1           40      Liver 7.764087 0.3503
## 2           30    unknown 5.545777 0.1596
## 3           20    unknown 1.109155 0.0000
## 4           30 Peritoneum 1.957439 0.1020
```

```
## 5           10     Pelvis 4.436621 0.4196
## 6           10    unknown 0.000000 0.0295
```

```
# check wheather all catagorical variables have been factorized
str(cleaned_data)
```

```
## 'data.frame':    7993 obs. of  15 variables:
##  $ Study               : Factor w/ 8 levels "Bladder Urothelial Carcinoma
TCGA/Firehose Legacy",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ PATIENT_ID          : Factor w/ 7309 levels "C3L-00104","C3L-00365",..:
333 334 335 335 336 337 338 339 339 340 ...
##  $ SAMPLE_ID           : Factor w/ 7612 levels "C3L-00104","C3L-00365",..:
213 214 215 216 217 218 219 220 221 222 ...
##  $ CANCER_TYPE         : Factor w/ 36 levels "Adrenal Gland",..: 8 22 13
13 13 22 23 18 18 8 ...
##  $ SITE_OF_TUMOR_TISSUE: Factor w/ 126 levels "Abdomen","Abdominal
Wall",..: 18 85 124 124 124 61 61 59 59 18 ...
##  $ SEX                 : Factor w/ 3 levels "Female","Male",..: 1 2 1 1 1
1 1 2 2 1 ...
##  $ RACE                : Factor w/ 5 levels "Asian","Black",..: 4 4 4 4 4
4 4 4 4 4 ...
##  $ AGE                 : num  46 47 47 33 33 47 47 33 46 47 ...
##  $ OS_STATUS           : Factor w/ 3 levels "0","1","unknown": 2 2 1 1 1 2
1 1 1 2 ...
##  $ OS_MONTHS           : num  71.1 8.71 8.81 8.81 71.1 ...
##  $ SMOKING_HISTORY     : Factor w/ 3 levels "Current","Former",..: 3 3 3 3
3 3 3 3 3 3 ...
##  $ TUMOR_PURITY        : int  40 30 20 30 10 10 30 90 90 30 ...
##  $ METASTATIC          : Factor w/ 112 levels "Abdomen","Abdominal
Wall",..: 43 107 107 75 72 107 107 43 107 107 ...
##  $ TMB                 : num  7.76 5.55 1.11 1.96 4.44 ...
##  $ FGA                 : num  0.35 0.16 0 0.102 0.42 ...
```

```
# check missing values again
missing_values <- colSums(is.na(cleaned_data))
print(missing_values)
```

```
##               Study          PATIENT_ID           SAMPLE_ID
##                   0                   0                   0
##         CANCER_TYPE SITE_OF_TUMOR_TISSUE                 SEX
##                   0                   0                   0
##                RACE                 AGE           OS_STATUS
##                   0                   0                   0
##           OS_MONTHS      SMOKING_HISTORY         TUMOR_PURITY
##                   0                   0                   0
##          METASTATIC                 TMB                 FGA
##                   0                   0                   0
```

There is no missing value after imputation.

```r
# Dataset for  automatic integration of fixed effects with mitml
# Extract all imputed datasets and convert to long format
imputed_long <- complete(imputed_data, action = "long", include = TRUE) ## 包含所有插补sets的总dataset

# factorizaion
imputed_long <- imputed_long %>%
  mutate(across(where(is.character), as.factor))

# restore data back into "mids" object
new_imputed_data <- mice::as.mids(imputed_long)

#  “mids” to  mitml.list
imputed_list <- mids2mitml.list(new_imputed_data)
```

Modeling with Mixed Effect Regression

In the modeling, we aimed to separately construct mixed-effects regression models for two measurements of Tumor Genomic Alterations—Tumor Mutation Burden (TMB) and Fraction Genome Altered (FGA)—to assess theImpact of various clinical factors on genomic alterations. We selected this model because we assumed that the clinical data exhibit a potential hierarchical structure (e.g., study and cancer type, study and patient). Additionally, the effects of some categorical variables, such as different data sources, are considered to be random, and there may be repeated or multiple records for the same patient.

Multicollinearity Check

```r
# Measure : VIF value
linear_model_TMB <- lm(TMB ~ AGE + SEX + RACE  +
TUMOR_PURITY+OS_STATUS+CANCER_TYPE +METASTATIC+SMOKING_HISTORY, data =
cleaned_data)
vif_values_TMB <- vif(linear_model_TMB)
print(vif_values_TMB)
```

```
##                       GVIF  Df GVIF^(1/(2*Df))
## AGE                1.529500   1        1.236729
## SEX                1.373927   2        1.082657
## RACE              14.387591   4        1.395560
## TUMOR_PURITY       1.525309   1        1.235034
## OS_STATUS          1.315191   2        1.070896
## CANCER_TYPE      101.642639  35        1.068249
## METASTATIC       118.208048 111        1.021730
## SMOKING_HISTORY    1.874670   2        1.170122
```

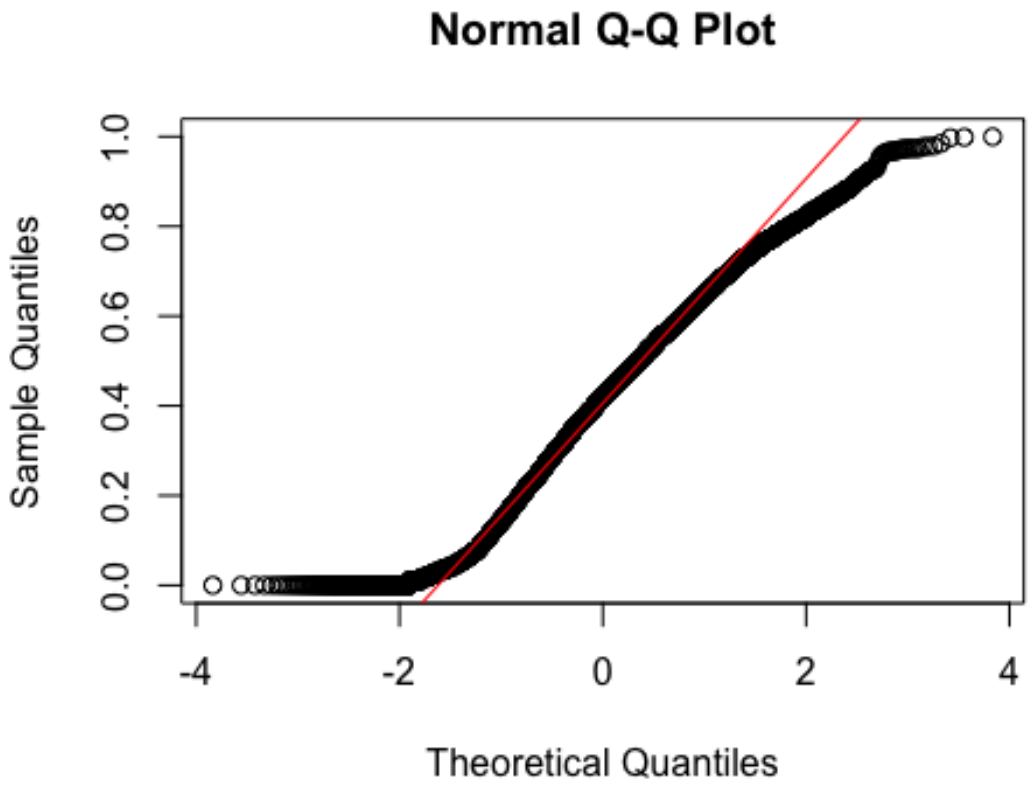
```r
# drop Metastatic due to multicolinear with cancer types
linear_model_FGA <- lm(TMB ~ AGE + SEX + RACE  +
TUMOR_PURITY+OS_STATUS+CANCER_TYPE +SMOKING_HISTORY, data = cleaned_data)
vif_values_FGA <- vif(linear_model_TMB)
print(vif_values_TMB)
```

```
##                       GVIF  Df GVIF^(1/(2*Df))
## AGE                1.529500   1        1.236729
## SEX                1.373927   2        1.082657
## RACE              14.387591   4        1.395560
## TUMOR_PURITY       1.525309   1        1.235034
## OS_STATUS          1.315191   2        1.070896
## CANCER_TYPE      101.642639  35        1.068249
## METASTATIC       118.208048 111        1.021730
## SMOKING_HISTORY    1.874670   2        1.170122
```

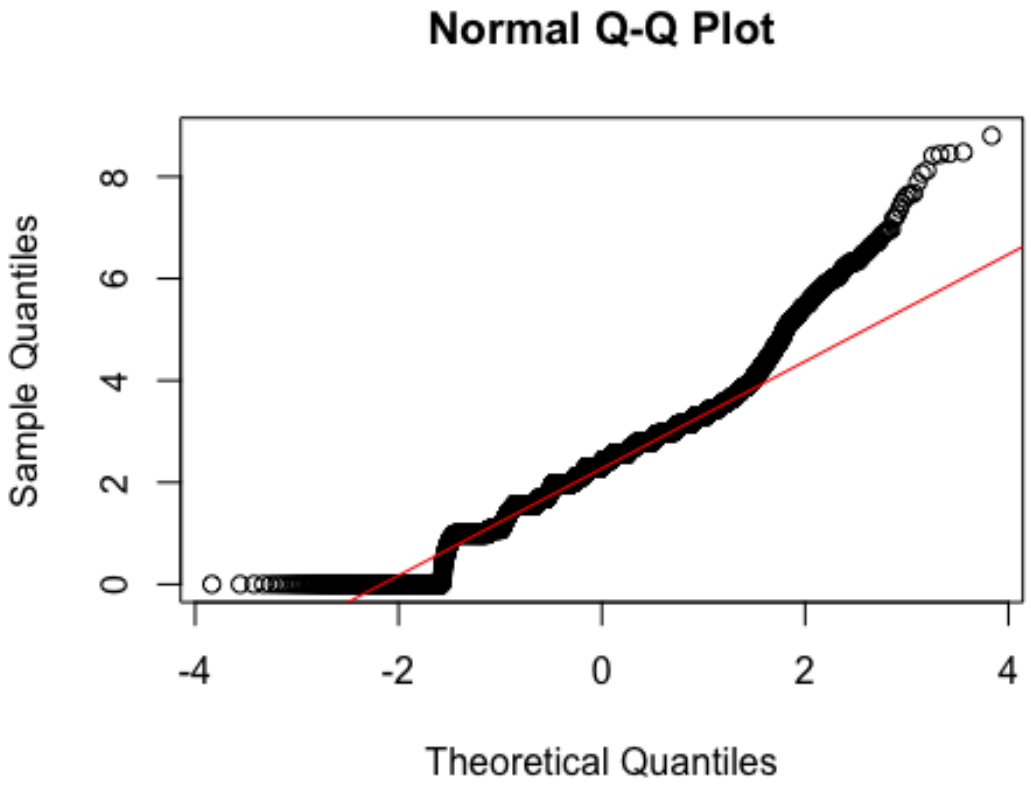
Before constructing the final models, we used a simple linear regression model to examine whether there was multicollinearity among the potential predictors. Variables showing multicollinearity were excluded from the subsequent modeling process due to the violation of the independence assumption required for reliable model estimation. Based on the VIF (Variance Inflation Factor) values, we identified a strong multicollinearity between the variables METASTATIC and Cancer Type. As a result, METASTATIC was excluded from the final model to maintain the independence assumption and ensure reliable model estimates.

Variable Distribution Visualization (Assumption check)

```
# hist(sqrt(cleaned_data$FGA), probability = TRUE, main = "Histogram with
Normal Curve For sqrt FGA")
qqnorm(sqrt(cleaned_data$FGA))
qqline(sqrt(cleaned_data$FGA), col = "red")
```



```r
# hist(log2(cleaned_data$TMB+1), probability = TRUE, main = "Histogram with
Normal Curve For transformed TMB")
qqnorm(log2(cleaned_data$TMB+1))
qqline(log2(cleaned_data$TMB+1), col = "red")
```



Since TMB and FGA did not follow a normal distribution, we applied transformations to meet the normality assumption for the response variables. The transformed TMB was calculated as log2(TMB + 1), and the transformed FGA was computed as sqrt(FGA). After these transformations, the response variables approximated a normal distribution. The Q-Q plots for transformed TMB and transformed FGA are shown above, illustrating their improved normality after transformation.

Modeling Explaination

Regarding the hierarchical structure of the random effects is not entirely clear due to the nature of the data. The dataset includes a PanCancer study covering multiple cancer types, as well as several specialized studies focused on one or two cancer types. Additionally, some studies from the same institution (MSK) share clinical data from overlapping patient cohorts. Therefore, in addition to modeling with crossed random effects, we also explored various alternative nested structures for the random effects. We then compared the model fit results across these different structures, aiming to investigate whether more complex model structures could improve model fit or enhance interpretability. Given that smoking

history was included as a fixed effect, we also attempted to build an additional model including its interaction with cancer type to explore whether smoking history has a stronger effect on TMB or FGA in certain cancer types, such as lung cancer. After completing the modeling process, we performed a normality test on the residuals. Since the database size exceeded the sample size limits of the Shapiro-Wilk test in the stats package, we assessed normality using a Q-Q plot of the residuals. Finally, we diagnosed and compared models' goodness of fit using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the R-squared calculation function from the performance package.

Modeling with TMB

```
# Mixed Effect Model for TMB

# main separated models
mixed_model_TMB_x_1 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_1)
mixed_model_TMB_x_2 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_2)
mixed_model_TMB_x_3 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_3)
mixed_model_TMB_x_4 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_4)
mixed_model_TMB_x_5 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_5)

# main integrated model
fit <- with(imputed_list, lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE)))
pooled_results_TMB_x <- testEstimates(fit)


# alternative model
mixed_model_TMB_in_1 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study:CANCER_TYPE )
, data = cleaned_data_1)
mixed_model_TMB_in_2 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study:CANCER_TYPE )
, data = cleaned_data_2)
mixed_model_TMB_in_3 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study:CANCER_TYPE )
, data = cleaned_data_3)
mixed_model_TMB_in_4 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study:CANCER_TYPE )
```

```
, data = cleaned_data_4)
mixed_model_TMB_in_5 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study:CANCER_TYPE )
, data = cleaned_data_5)
```

We used the five imputed datasets to model the transformed TMB. For the fixed effects, we integrated the results using the testEstimates function from the mitml package. However, this function does not support the integration of random effects and does not calculate AIC, BIC, or R-squared. To address this limitation, we manually modeled the transformed TMB on each imputed dataset, enabling us to capture random effects and perform model evaluation and comparison across datasets.

Diagnosis and Comparison for TMB Models

```
# Main models for TMB
aic_values <- c()
bic_values <- c()
r2_values <- c()


model_list <- list(mixed_model_TMB_x_1, mixed_model_TMB_x_2,
mixed_model_TMB_x_3, mixed_model_TMB_x_4, mixed_model_TMB_x_5)

for (model in model_list) {
  aic_values <- c(aic_values, AIC(model))
  bic_values <- c(bic_values, BIC(model))
  r2_values <- c(r2_values, performance::r2(model)$R2_conditional)  # For
conditional R2
}

avg_aic <- mean(aic_values)
avg_bic <- mean(bic_values)
avg_r2 <- mean(r2_values)

cat("Diagnosis for TMB main model \n")

## Diagnosis for TMB main model

cat("Average AIC:", avg_aic, "\n")

## Average AIC: 22511.11

cat("Average BIC:", avg_bic, "\n")

## Average BIC: 22615.9

cat("Average R-squared:", avg_r2, "\n")

## Average R-squared: 0.8532057

cat("\n")
```

```r
# Alter Models for TMB
aic_values <- c()
bic_values <- c()
r2_values <- c()


model_list <- list(mixed_model_TMB_in_1, mixed_model_TMB_in_2,
mixed_model_TMB_in_3, mixed_model_TMB_in_4, mixed_model_TMB_in_5)

for (model in model_list) {
  aic_values <- c(aic_values, AIC(model))
  bic_values <- c(bic_values, BIC(model))
  r2_values <- c(r2_values, performance::r2(model)$R2_conditional)  # For
conditional R2
}

avg_aic <- mean(aic_values)
avg_bic <- mean(bic_values)
avg_r2 <- mean(r2_values)

cat("Diagnosis for TMB alternative model \n")

## Diagnosis for TMB alternative model

cat("Average AIC:", avg_aic, "\n")

## Average AIC: 22512.74

cat("Average BIC:", avg_bic, "\n")

## Average BIC: 22610.55

cat("Average R-squared:", avg_r2, "\n")

## Average R-squared: 0.8360782
```

From the model diagnostics, we found that more complex alternative models incorporating interactions did not show any advantage in terms of goodness of fit compared to the main model with crossed random effects. Despite adding interaction terms, these alternative models did not outperform the simpler structure of the main model.

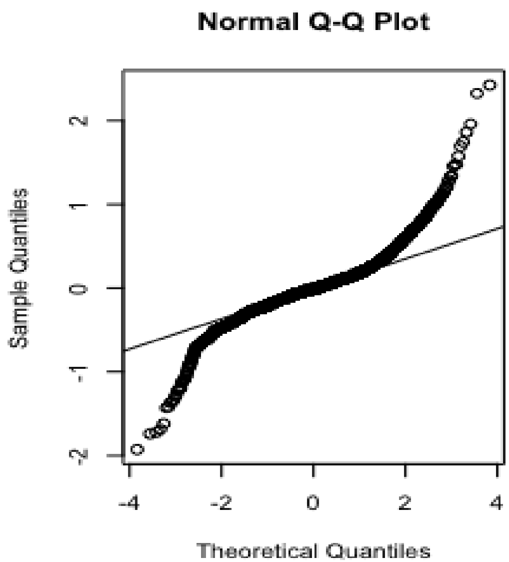
```r
anova(mixed_model_TMB_x_1,mixed_model_TMB_in_1)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_1
## Models:
## mixed_model_TMB_in_1: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study:CANCER_TYPE)
## mixed_model_TMB_x_1: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
##                       npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
```

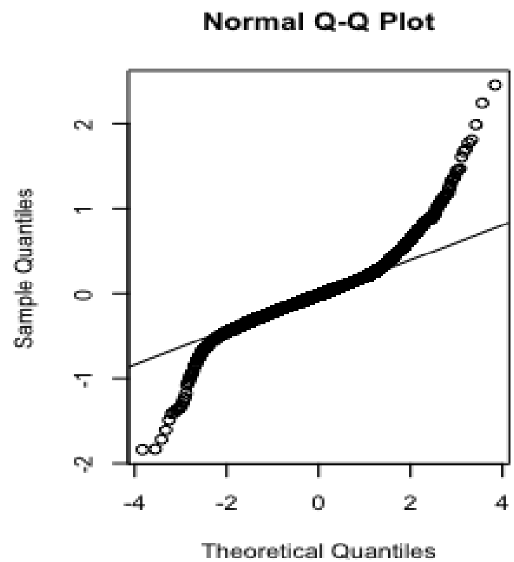
```
## mixed_model_TMB_in_1   14 22752 22849 -11362    22724
## mixed_model_TMB_x_1    15 22753 22858 -11361    22723 0.8148  1     0.3667

anova(mixed_model_TMB_x_2,mixed_model_TMB_in_2)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_2
## Models:
## mixed_model_TMB_in_2: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study:CANCER_TYPE)
## mixed_model_TMB_x_2: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
##                       npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
## mixed_model_TMB_in_2   14 22388 22486 -11180    22360
## mixed_model_TMB_x_2    15 22387 22492 -11179    22357 2.647  1     0.1037

anova(mixed_model_TMB_x_3,mixed_model_TMB_in_3)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_3
## Models:
## mixed_model_TMB_in_3: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study:CANCER_TYPE)
## mixed_model_TMB_x_3: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
##                       npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## mixed_model_TMB_in_3   14 22713 22810 -11342    22685
## mixed_model_TMB_x_3    15 22713 22818 -11341    22683 1.8439  1     0.1745

anova(mixed_model_TMB_x_4,mixed_model_TMB_in_4)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_4
## Models:
## mixed_model_TMB_in_4: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study:CANCER_TYPE)
## mixed_model_TMB_x_4: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
##                       npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## mixed_model_TMB_in_4   14 21727 21825 -10850    21699
## mixed_model_TMB_x_4    15 21726 21830 -10848    21696 3.6185  1    0.05714
.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(mixed_model_TMB_x_5,mixed_model_TMB_in_5)

## refitting model(s) with ML (instead of REML)
```

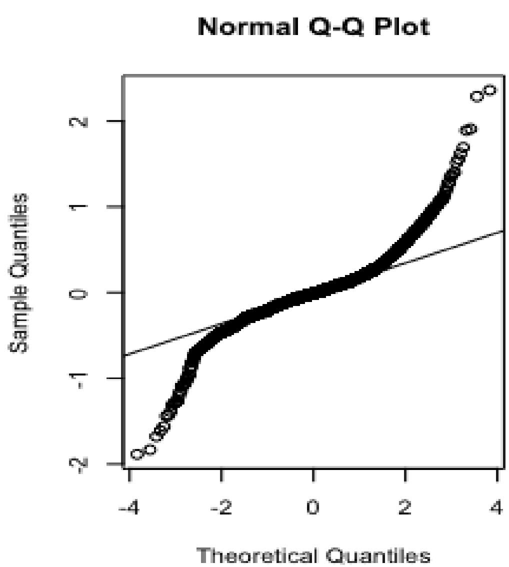
```
## Data: cleaned_data_5
## Models:
## mixed_model_TMB_in_5: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study:CANCER_TYPE)
## mixed_model_TMB_x_5: log2(TMB + 1) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
##                       npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## mixed_model_TMB_in_5    14 22711 22809 -11342    22683
## mixed_model_TMB_x_5     15 22711 22816 -11341    22681 2.1184  1     0.1455
```

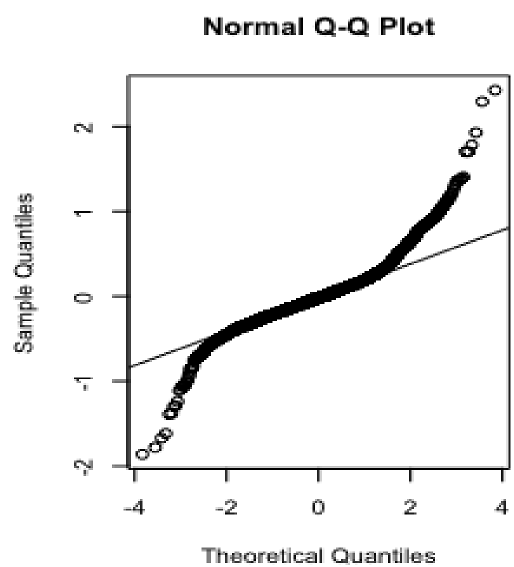
Additionally, ANOVA tests comparing the two models across all imputed data sets returned p-values greater than 0.05, indicating that the differences between the models were not statistically significant. This further confirms that the more complex models with interactions do not offer significant improvements over the main model.

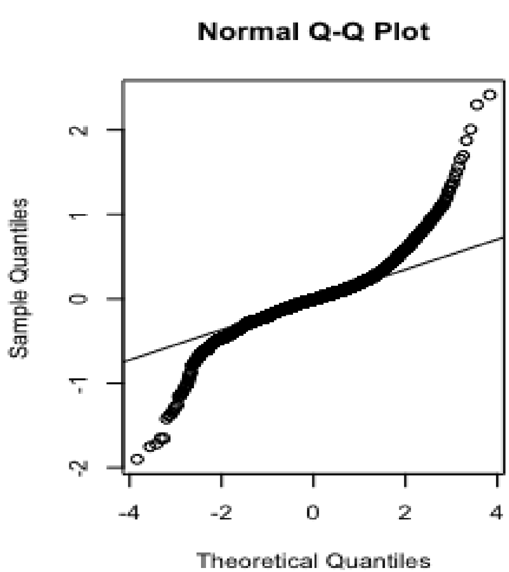
```
# check normality of residuals
par(mfrow = c(3,2))
qqnorm(residuals(mixed_model_TMB_x_1))
qqline(residuals(mixed_model_TMB_x_1))
qqnorm(residuals(mixed_model_TMB_x_2))
qqline(residuals(mixed_model_TMB_x_2))
qqnorm(residuals(mixed_model_TMB_x_3))
qqline(residuals(mixed_model_TMB_x_3))
qqnorm(residuals(mixed_model_TMB_x_4))
qqline(residuals(mixed_model_TMB_x_4))
qqnorm(residuals(mixed_model_TMB_x_5))
qqline(residuals(mixed_model_TMB_x_5))

# # scatter plot for residuals vs fitted values
# par(mfrow = c(3,2))
# plot(fitted(mixed_model_TMB_x_1), residuals(mixed_model_TMB_x_1))
# abline(h = 0, col = "blue")
# plot(fitted(mixed_model_TMB_x_2), residuals(mixed_model_TMB_x_2))
# abline(h = 0, col = "blue")
# plot(fitted(mixed_model_TMB_x_3), residuals(mixed_model_TMB_x_3))
# abline(h = 0, col = "blue")
# plot(fitted(mixed_model_TMB_x_4), residuals(mixed_model_TMB_x_4))
# abline(h = 0, col = "blue")
# plot(fitted(mixed_model_TMB_x_5), residuals(mixed_model_TMB_x_5))
# abline(h = 0, col = "blue")
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-4
-2
0
2
4


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-4
-2
0
2
4


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-4
-2
0
2
4


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-4
-2
0
2
4


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-4
-2
0
2
4

The Q-Q plots for the residuals of the fitted models across each data set are shown above. The residuals generally follow a normal distribution, though there is some deviation at the tails. This indicates that while the models perform well in most of the distribution, there may still be some bias in the extreme values.

Modeling with FGA

```
# Mixed Effect Model for FGA

# main separated models
mixed_model_FGA_x_1 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_1)
mixed_model_FGA_x_2 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_2)
mixed_model_FGA_x_3 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_3)
mixed_model_FGA_x_4 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_4)
mixed_model_FGA_x_5 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study)
+(1|CANCER_TYPE) , data = cleaned_data_5)

# main integrated model
fit <- with(imputed_list, lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +
(1|CANCER_TYPE)))
pooled_results_FGA_x <- testEstimates(fit)


# alternative model
mixed_model_FGA_n_1 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study) +
(1|Study:CANCER_TYPE) , data = cleaned_data_1)
mixed_model_FGA_n_2 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study) +
(1|Study:CANCER_TYPE) , data = cleaned_data_2)
mixed_model_FGA_n_3 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study) +
(1|Study:CANCER_TYPE) , data = cleaned_data_3)
mixed_model_FGA_n_4 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study) +
(1|Study:CANCER_TYPE) , data = cleaned_data_4)
mixed_model_FGA_n_5 <- lmerTest::lmer(sqrt(FGA) ~ AGE + SEX + RACE  +
TUMOR_PURITY + SMOKING_HISTORY  +  (1 | PATIENT_ID ) + (1 | Study) +
(1|Study:CANCER_TYPE) , data = cleaned_data_5)
```

We used the five imputed datasets to model the transformed FGA. For the fixed effects, we integrated the results using the testEstimates function from the mitml package. However, this function does not support the integration of random effects and does not calculate AIC, BIC, or R-squared. To address this limitation, we manually modeled the transformed TMB on each imputed dataset, enabling us to capture random effects and perform model evaluation and comparison across datasets.

Diagnosis and Comparison for FGA Models

```
# Main models for FGA
aic_values <- c()
bic_values <- c()
r2_values <- c()


model_list <- list(mixed_model_FGA_x_1, mixed_model_FGA_x_2,
mixed_model_FGA_x_3, mixed_model_FGA_x_4, mixed_model_FGA_x_5)

for (model in model_list) {
  aic_values <- c(aic_values, AIC(model))
  bic_values <- c(bic_values, BIC(model))
  r2_values <- c(r2_values, performance::r2(model)$R2_conditional)  # For
conditional R2
}

avg_aic <- mean(aic_values)
avg_bic <- mean(bic_values)
avg_r2 <- mean(r2_values)

cat("Diagnosis for FGA main model \n")

## Diagnosis for FGA main model

cat("Average AIC:", avg_aic, "\n")

## Average AIC: -2965.401

cat("Average BIC:", avg_bic, "\n")

## Average BIC: -2860.606

cat("Average R-squared:", avg_r2, "\n")

## Average R-squared: 0.7931746

cat("\n")

# Alter Models for FGA
aic_values <- c()
bic_values <- c()
r2_values <- c()
```

```r
model_list <- list(mixed_model_FGA_n_1, mixed_model_FGA_n_2,
mixed_model_FGA_n_3, mixed_model_FGA_n_4, mixed_model_FGA_n_5)

for (model in model_list) {
  aic_values <- c(aic_values, AIC(model))
  bic_values <- c(bic_values, BIC(model))
  r2_values <- c(r2_values, performance::r2(model)$R2_conditional)  # For
conditional R2
}

avg_aic <- mean(aic_values)
avg_bic <- mean(bic_values)
avg_r2 <- mean(r2_values)

cat("Diagnosis for FGA alternative model \n")
```

```
## Diagnosis for FGA alternative model
```

```r
cat("Average AIC:", avg_aic, "\n")
```

```
## Average AIC: -2967.029
```

```r
cat("Average BIC:", avg_bic, "\n")
```

```
## Average BIC: -2862.234
```

```r
cat("Average R-squared:", avg_r2, "\n")
```

```
## Average R-squared: 0.7783493
```

From the model diagnostics, we found that more complex alternative models incorporating interactions did not show any advantage in terms of goodness of fit compared to the main model with crossed random effects. Despite adding interaction terms, these alternative models did not outperform the simpler structure of the main model.

```r
anova(mixed_model_FGA_x_1,mixed_model_FGA_n_1)
```

```
## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_1
## Models:
## mixed_model_FGA_x_1: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
## mixed_model_FGA_n_1: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | Study:CANCER_TYPE)
##                     npar     AIC     BIC logLik deviance  Chisq Df
Pr(>Chisq)
## mixed_model_FGA_x_1   15 -2565.4 -2460.6 1297.7  -2595.4
## mixed_model_FGA_n_1   15 -2567.2 -2462.4 1298.6  -2597.2 1.8067  0
```

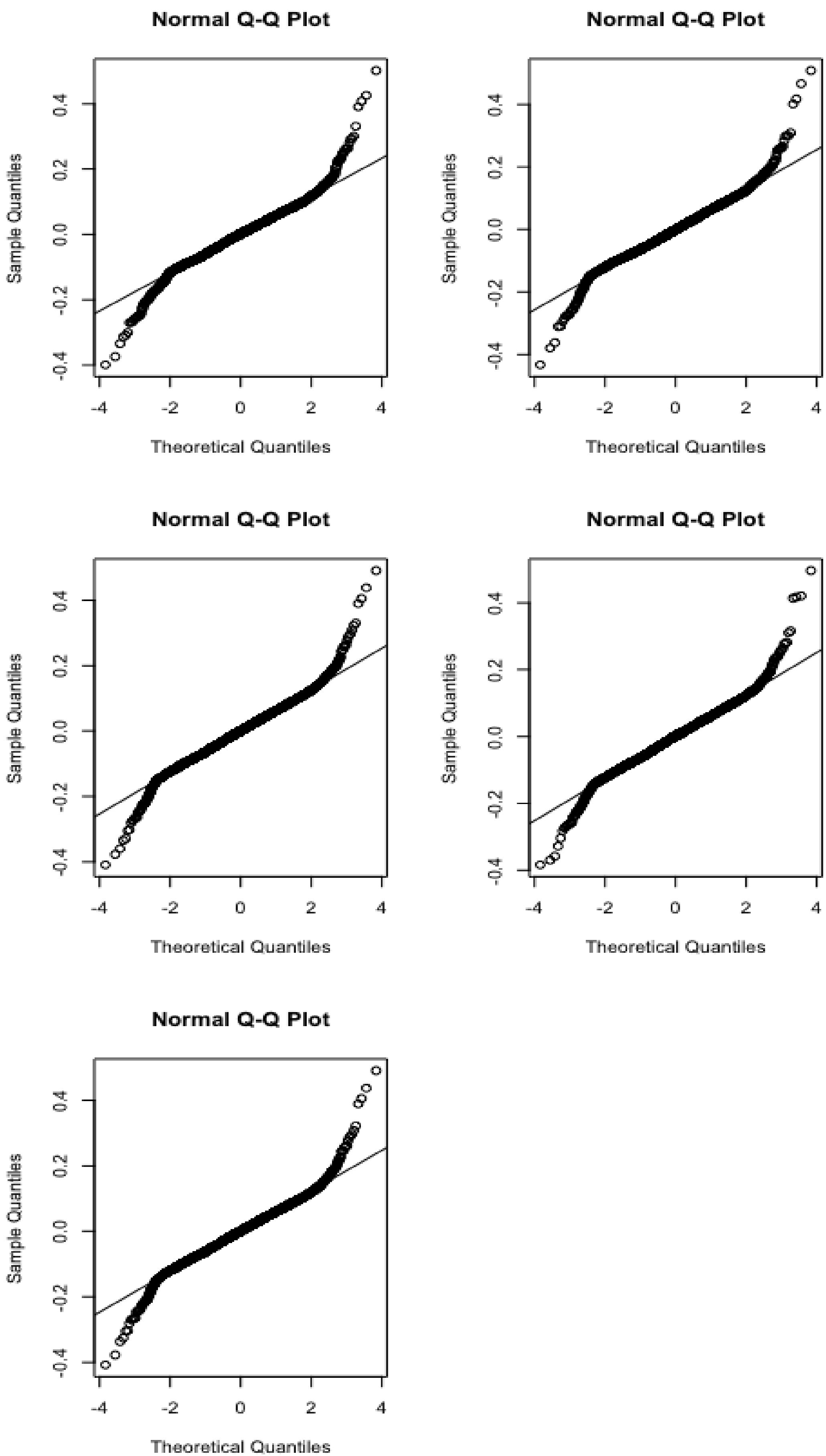
```r
anova(mixed_model_FGA_x_2,mixed_model_FGA_n_2)
```

```
## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_2
## Models:
## mixed_model_FGA_x_2: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
## mixed_model_FGA_n_2: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | Study:CANCER_TYPE)
##                     npar     AIC     BIC logLik deviance  Chisq Df
Pr(>Chisq)
## mixed_model_FGA_x_2   15 -3405.1 -3300.3 1717.6  -3435.1
## mixed_model_FGA_n_2   15 -3405.2 -3300.4 1717.6  -3435.2 0.0426  0

anova(mixed_model_FGA_x_3,mixed_model_FGA_n_3)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_3
## Models:
## mixed_model_FGA_x_3: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
## mixed_model_FGA_n_3: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | Study:CANCER_TYPE)
##                     npar     AIC     BIC logLik deviance  Chisq Df
Pr(>Chisq)
## mixed_model_FGA_x_3   15 -3088.6 -2983.8 1559.3  -3118.6
## mixed_model_FGA_n_3   15 -3092.9 -2988.1 1561.5  -3122.9 4.3404  0

anova(mixed_model_FGA_x_4,mixed_model_FGA_n_4)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_4
## Models:
## mixed_model_FGA_x_4: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
## mixed_model_FGA_n_4: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | Study:CANCER_TYPE)
##                     npar     AIC     BIC logLik deviance  Chisq Df
Pr(>Chisq)
## mixed_model_FGA_x_4   15 -3222.1 -3117.3   1626  -3252.1
## mixed_model_FGA_n_4   15 -3226.1 -3121.3   1628  -3256.1 4.0234  0

anova(mixed_model_FGA_x_5,mixed_model_FGA_n_5)

## refitting model(s) with ML (instead of REML)

## Data: cleaned_data_5
## Models:
## mixed_model_FGA_x_5: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | CANCER_TYPE)
## mixed_model_FGA_n_5: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR_PURITY +
```

```
SMOKING_HISTORY + (1 | PATIENT_ID) + (1 | Study) + (1 | Study:CANCER_TYPE)
##                      npar     AIC     BIC logLik deviance  Chisq Df
Pr(>Chisq)
## mixed_model_FGA_x_5   15 -2989.3 -2884.5 1509.6  -3019.3
## mixed_model_FGA_n_5   15 -2993.0 -2888.2 1511.5  -3023.0 3.7088  0
```

Additionally, ANOVA tests comparing the two models across all imputed data sets indicated that the differences between the models were not statistically significant. This further confirms that the more complex models with interactions do not offer significant improvements over the main model.

```
# check normality of residuals
par(mfrow = c(3,2))
qqnorm(residuals(mixed_model_FGA_x_1))
qqline(residuals(mixed_model_FGA_x_1))
qqnorm(residuals(mixed_model_FGA_x_2))
qqline(residuals(mixed_model_FGA_x_2))
qqnorm(residuals(mixed_model_FGA_x_3))
qqline(residuals(mixed_model_FGA_x_3))
qqnorm(residuals(mixed_model_FGA_x_4))
qqline(residuals(mixed_model_FGA_x_4))
qqnorm(residuals(mixed_model_FGA_x_5))
qqline(residuals(mixed_model_FGA_x_5))
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

The Q-Q plots for the residuals of the fitted models across each data set are shown above. The residuals generally follow a normal distribution, though there is some deviation at the tails. This indicates that while the models perform well in most of the distribution, there may still be some bias in the extreme values.

Smoking History v.s. Cancer Type

```
# main separated models

SH_CT_TMB_1 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE) , data = cleaned_data_1)
SH_CT_TMB_2 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE) , data = cleaned_data_2)
SH_CT_TMB_3 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE) , data = cleaned_data_3)
SH_CT_TMB_4 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE) , data = cleaned_data_4)
SH_CT_TMB_5 <- lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE) , data = cleaned_data_5)

fit_cs <- with(imputed_list, lmerTest::lmer(log2(TMB+1) ~ AGE + SEX + RACE  +
TUMOR_PURITY+SMOKING_HISTORY + (1 | PATIENT_ID ) +(1 | Study) +(1 +
SMOKING_HISTORY|CANCER_TYPE)))
pooled_results_cancer_smoke <- testEstimates(fit_cs)
```

We also constructed an additional model to explore whether the effect of smoking history on TMB varies across different cancer types. In this model, smoking history was included as a random slope to account for its varying influence across cancer types. The mitml package was used to integrate the fixed effects, while manual modeling was employed to capture the random effects from each dataset's fit.

Results

Result from Main Model with Transformed TMB

Our Main Model with Transformed TMB is :

$$\log_2(\text{TMB}+1) = \beta_0 + \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{SEX} + \beta_3 \cdot \text{RACE} + \beta_4 \cdot \text{TUMOR\_PURITY} + \beta_5 \cdot \text{SMOKING\_HISTORY} + u_{\text{PATIENT\_ID}} + u_{\text{Study}} + u_{\text{CANCER\_TYPE}} + \epsilon$$

```
# Extracting fixed effects
summary(pooled_results_TMB_x)
```

```
##
## Call:
##
## testEstimates(model = fit)
##
## Final parameter estimates and inferences obtained from 5 imputed data
sets.
##
##                        Estimate Std.Error   t.value        df   P(>|t|)
RIV       FMI
## (Intercept)               1.908     0.946     2.017 4.519e+00     0.106
15.903     0.957
## AGE                      -0.005     0.017    -0.286 4.030e+00     0.789
270.155     0.997
## SEXMale                   0.075     0.029     2.545 1.514e+02     0.012
0.194     0.173
## SEXUnknown                1.751     0.308     5.676 6.372e+02     0.000
0.086     0.082
## RACEBlack                 0.024     0.103     0.232 4.435e+05     0.817
0.003     0.003
## RACEOther                 0.023     0.112     0.202 1.218e+04     0.840
0.018     0.018
## RACEUnknown               0.140     0.159     0.880 6.210e+03     0.379
0.026     0.026
## RACEWhite                 0.075     0.085     0.875 8.098e+01     0.384
0.286     0.241
## TUMOR_PURITY              0.004     0.001     2.920 5.466e+00     0.030
5.920     0.890
## SMOKING_HISTORYFormer    -0.143     0.062    -2.316 6.586e+01     0.024
0.327     0.268
## SMOKING_HISTORYNever     -0.320     0.064    -4.969 9.490e+01     0.000
0.258     0.222
##
## Unadjusted hypothesis test as appropriate in larger samples.
```

From the integrated fixed effect we can find:

- The estimate of Age is -0.005 with a p-value of 0.789, indicating no significant linear relationship between age and TMB. While the negative coefficient suggests a slight decrease in TMB with age, this effect is not statistically significant.

- The estimate for males is 0.075, with a p-value of 0.012, suggesting that males have significantly higher TMB compared to females.
- The estimate for unknown sex is 1.751, with a p-value of 0.000, indicating a significantly higher TMB for individuals with unknown sex compared to females.

- For the different race categories (Black, Other, Unknown, White), the estimates and p-values indicate that race does not have a significant effect

on TMB. All p-values are greater than 0.05, suggesting that race differences are not statistically significant in this model.

- The estimate for TUMOR_PURITY is 0.004 with a p-value of 0.030, indicating a significant positive relationship between tumor purity and TMB. As tumor purity increases, TMB significantly increases.

- The estimate for "Former smoker" is -0.143, with a p-value of 0.024, suggesting that former smokers have significantly lower TMB compared to current smokers.
- The estimate for "Never smoker" is -0.320, with a p-value of 0.000, indicating that individuals who have never smoked have significantly lower TMB compared to current smokers, with a more substantial reduction than former smokers.

The fixed effects in the model reveal that sex, tumor purity, and smoking history are significant predictors of TMB. However, age and race do not significantly impact TMB in this model.

```r
# Extracting random effect
## Extraction variance
variance_list <- list()

for (i in 1:5) {
  model_name <- get(paste0("mixed_model_TMB_x_", i))

  re_variance <- as.data.frame(VarCorr(model_name))

  re_variance_df <- data.frame(
    group = re_variance$grp,
    term = re_variance$var1,
    variance = re_variance$vcov
  )

  variance_list[[i]] <- rbind(re_variance_df)
}

variance_df <- do.call(rbind, variance_list)

# Variance of each variables (按 group 和 term 分组)
mean_variances <- aggregate(variance ~ group , data = variance_df, FUN =
mean)

print(mean_variances)
```

```
##         group  variance
## 1 CANCER_TYPE 0.2984433
## 2  PATIENT_ID 0.7995933
```

```
## 3    Residual 0.2339470
## 4       Study 0.1975187
```

From the integrated random effect we can find:

• The variance for cancer type is 0.2984, indicating that there is considerable variability in TMB between different cancer types. This means that part of the variability in TMB can be explained by differences across cancer types.

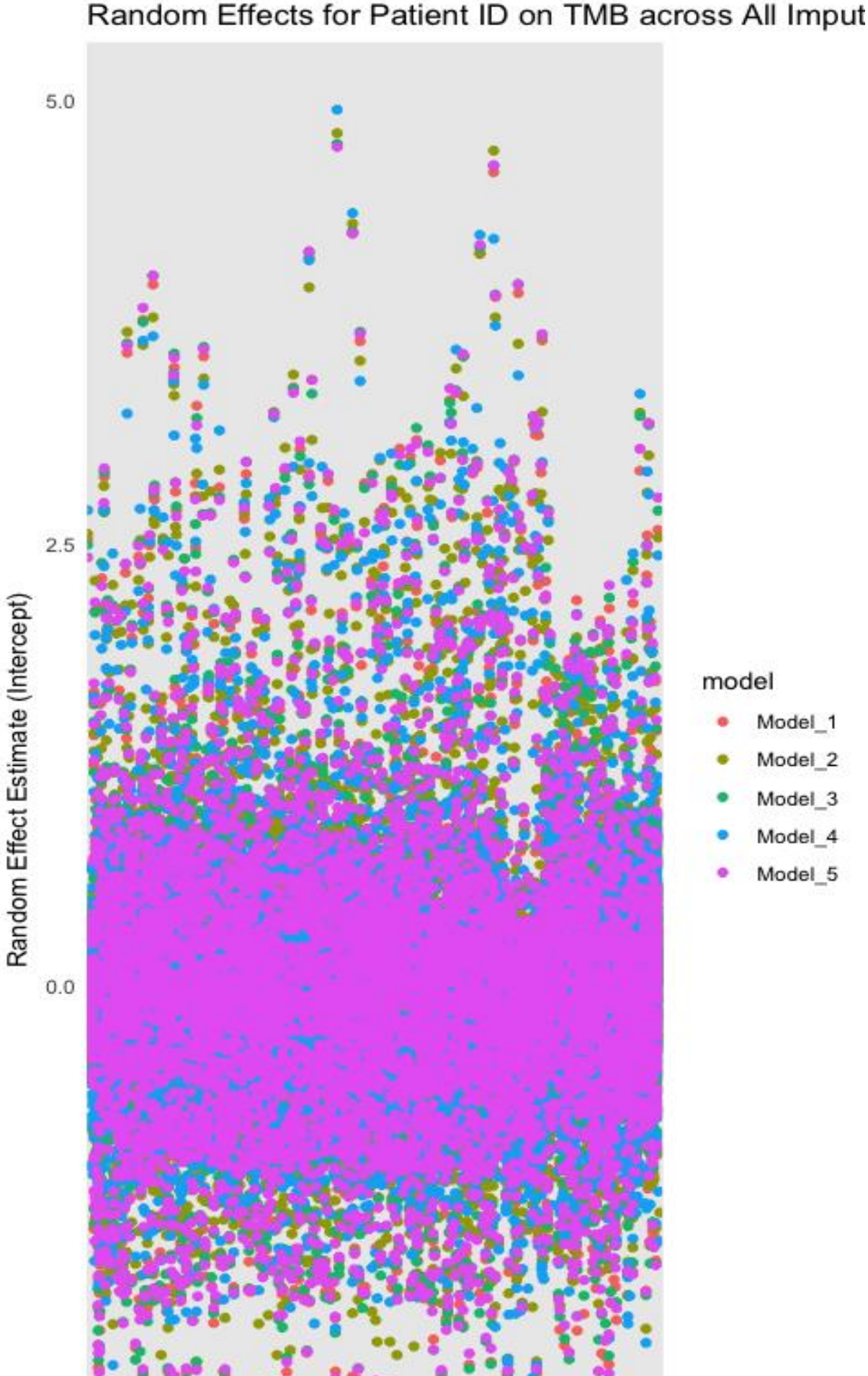
• The variance for patient ID is 0.7996, showing substantial variability in TMB between individual patients. This is the largest source of variance in the model, meaning that patient-level differences have the greatest influence on TMB.

• The variance for studies is 0.1975, suggesting that there are some differences in TMB across different studies, though the influence is smaller compared to patient-level variance.

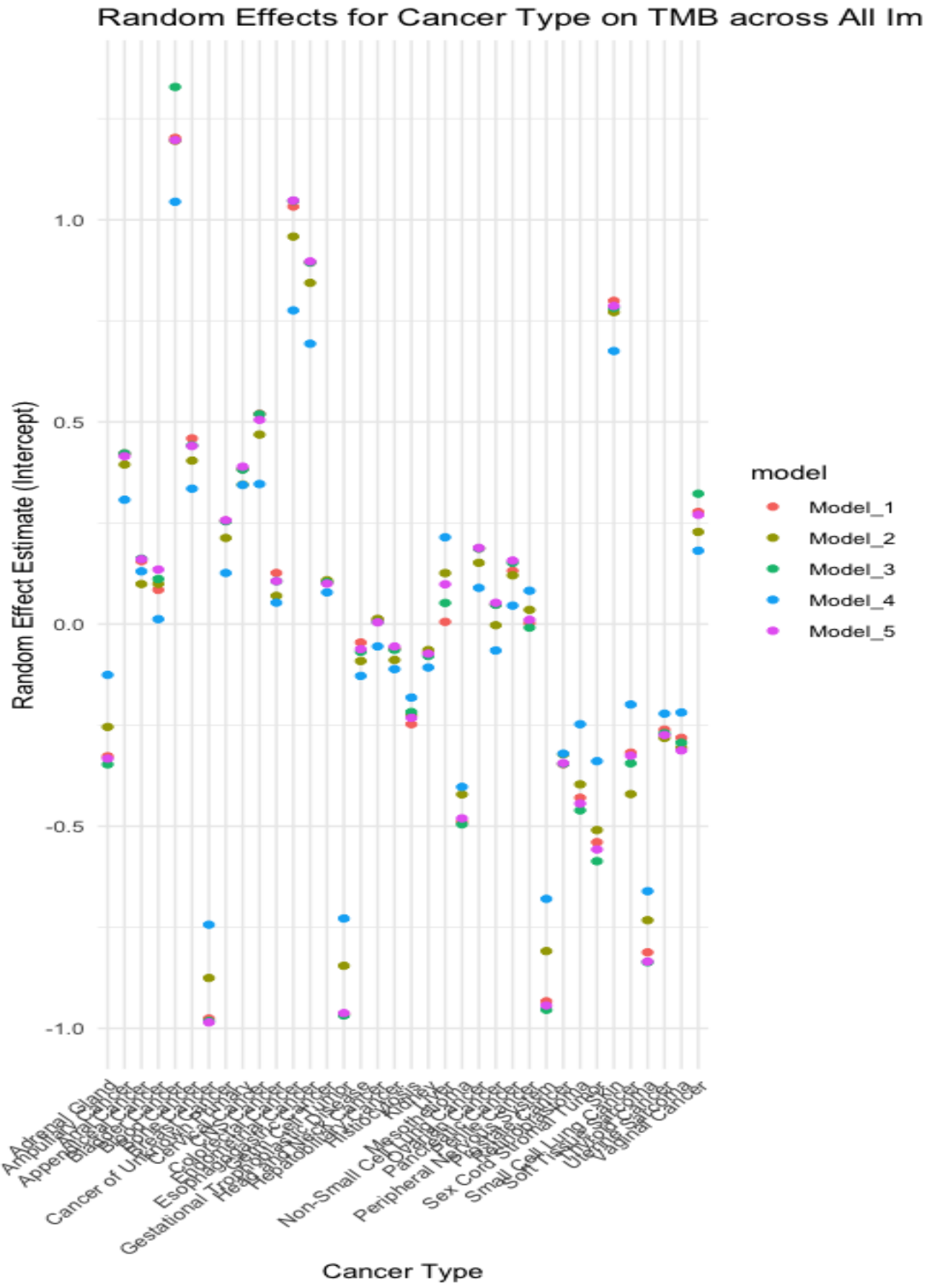
• The residual variance is 0.2339, representing the unexplained variation in TMB. Residual variance reflects random errors or other factors not included in the model.

```
## Extraction intercept (visualizaion all datasets' randam effect on a single plot)
### Extract the random effects of each model and add the model label
extract_random_effects <- function(model, model_name) {
  random_effects <- ranef(model)
  #  get Patient_ID, Cancer_Type, Study random effect
  patient_id_effects <- data.frame(PatientID = rownames(random_effects$PATIENT_ID),
                                   Effect = random_effects$PATIENT_ID$`(Intercept)`,
                                   group = "Patient_ID",
                                   model = model_name)

  cancer_type_effects <- data.frame(CancerType = rownames(random_effects$CANCER_TYPE),
                                    Effect = random_effects$CANCER_TYPE$`(Intercept)`,
                                    group = "Cancer_Type",
                                    model = model_name)

  study_effects <- data.frame(Study = rownames(random_effects$Study),
                              Effect = random_effects$Study$`(Intercept)`,
                              group = "Study",
                              model = model_name)

  return(list(patient_id = patient_id_effects, cancer_type = cancer_type_effects, study = study_effects))
}
```

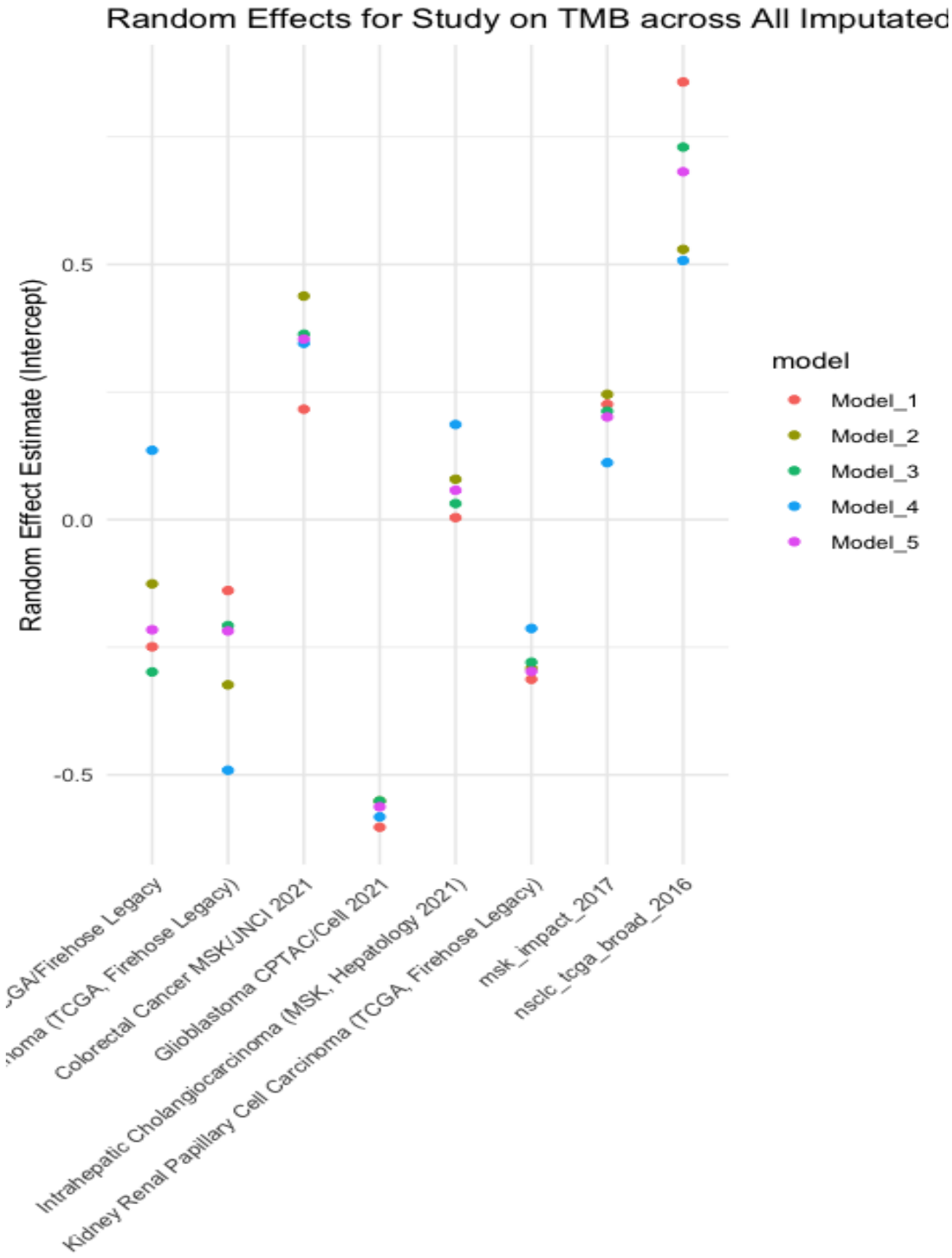
```r
# get 5 models‘ random eff
random_effects_1 <- extract_random_effects(mixed_model_TMB_x_1, "Model_1")
random_effects_2 <- extract_random_effects(mixed_model_TMB_x_2, "Model_2")
random_effects_3 <- extract_random_effects(mixed_model_TMB_x_3, "Model_3")
random_effects_4 <- extract_random_effects(mixed_model_TMB_x_4, "Model_4")
random_effects_5 <- extract_random_effects(mixed_model_TMB_x_5, "Model_5")

# put togehter by vairbale name
combined_patient_id <- rbind(random_effects_1$patient_id,
random_effects_2$patient_id, random_effects_3$patient_id,
                             random_effects_4$patient_id,
random_effects_5$patient_id)

combined_cancer_type <- rbind(random_effects_1$cancer_type,
random_effects_2$cancer_type, random_effects_3$cancer_type,
                              random_effects_4$cancer_type,
random_effects_5$cancer_type)

combined_study <- rbind(random_effects_1$study, random_effects_2$study,
random_effects_3$study,
                        random_effects_4$study, random_effects_5$study)

# Visual。。。。
ggplot(combined_patient_id, aes(x = PatientID, y = Effect, color = model)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Patient ID on TMB across All Imputated
Datasets",
       x = "Patient ID", y = "Random Effect Estimate (Intercept)") +
  # theme(axis.text.x = element_text(angle = 45, hjust = 1))
  theme(axis.text.x = element_blank())
```


Random Effects for Patient ID on TMB across All Imput
Random Effect Estimate (Intercept)
5.0
2.5
0.0
model
Model_1
Model_2
Model_3
Model_4
Model_5

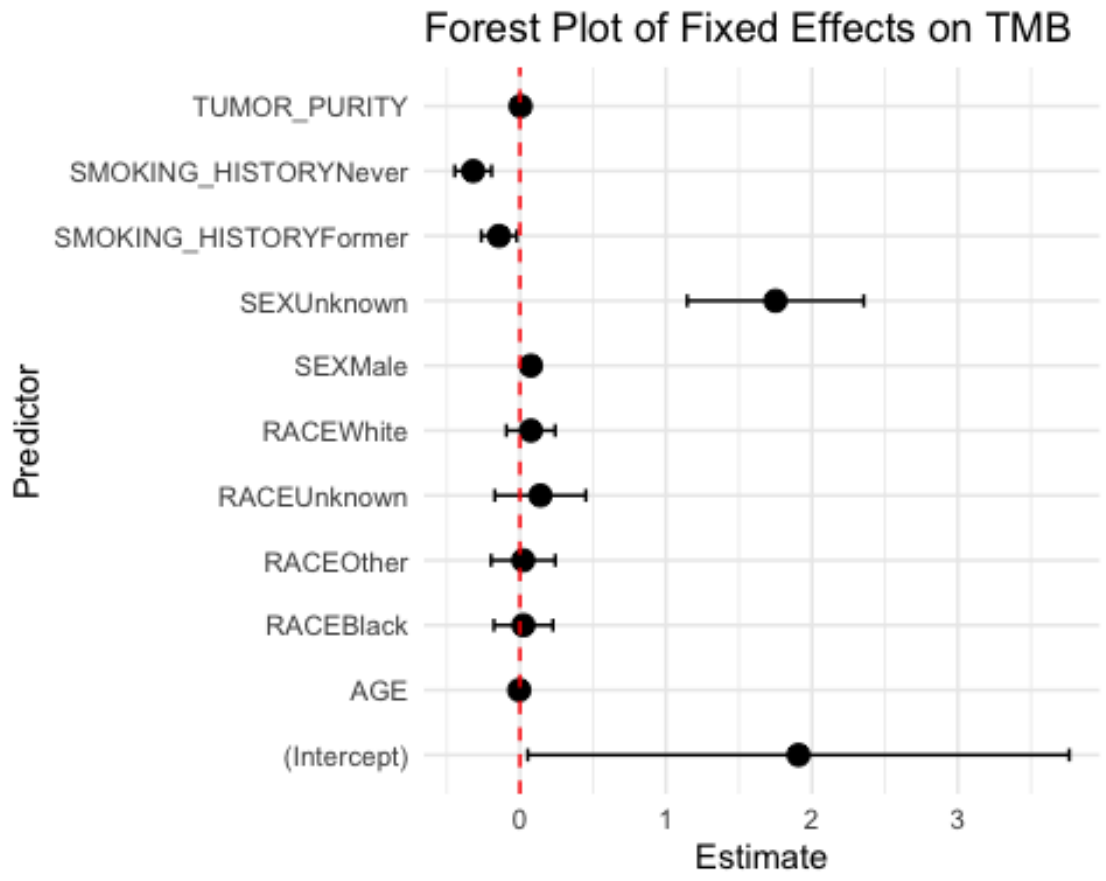
```r
ggplot(combined_cancer_type, aes(x = CancerType, y = Effect, color = model))
+
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Cancer Type on TMB across All Imputated
Datasets",
       x = "Cancer Type", y = "Random Effect Estimate (Intercept)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Random Effects for Cancer Type on TMB across All Im
Random Effect Estimate (Intercept)
1.0
0.5
0.0
-0.5
-1.0
Cancer Type
model
Model_1
Model_2
Model_3
Model_4
Model_5

```r
ggplot(combined_study, aes(x = Study, y = Effect, color = model)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Study on TMB across All Imputated
Datasets",
       x = "Study", y = "Random Effect Estimate (Intercept)") +
  scale_y_continuous(limits = c(min(combined_study$Effect),
max(combined_study$Effect))) +
  # theme(axis.text.x = element_blank())
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Random Effects for Study on TMB across All Imputated
Random Effect Estimate (Intercept)
0.5
0.0
-0.5
model
Model_1
Model_2
Model_3
Model_4
Model_5
GA/Firehose Legacy
noma (TCGA, Firehose Legacy)
Colorectal Cancer MSK/JNCI 2021
Glioblastoma CPTAC/Cell 2021
Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021)
Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy)
msk_impact_2017
nsclc_tcga_broad_2016

```r
# 所有插补集的fixed effect 的可视化
# 从 pooled_results 中提取固定效应估计
estimates <- pooled_results_TMB_x$estimates

forest_data <- data.frame(
  term = rownames(estimates),
  estimate = estimates[, "Estimate"],
  std.error = estimates[, "Std.Error"],
  conf.low = estimates[, "Estimate"] - 1.96 * estimates[, "Std.Error"],
  conf.high = estimates[, "Estimate"] + 1.96 * estimates[, "Std.Error"]
)


ggplot(forest_data, aes(x = estimate, y = term)) +
  geom_point(size = 3) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  theme_minimal() +
  labs(title = "Forest Plot of Fixed Effects on TMB", x = "Estimate", y =
"Predictor") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red")
```



Results from Main Model with Transformed FGA

Our Main Model with Transformed FGA is :

$$\sqrt{FGA} = \beta_0 + \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{SEX} + \beta_3 \cdot \text{RACE} + \beta_4 \cdot \text{TUMOR\_PURITY} + \beta_5 \cdot \text{SMOKING\_HISTORY} + u_{\text{PATIENT\_ID}} + u_{\text{Study}} + u_{\text{CANCER\_TYPE}} + \epsilon$$

```
# Extracting fixed effects
summary(pooled_results_FGA_x)

##
## Call:
##
## testEstimates(model = fit)
##
## Final parameter estimates and inferences obtained from 5 imputed data
sets.
##
##                          Estimate Std.Error   t.value        df   P(>|t|)
RIV       FMI
## (Intercept)               0.362     0.102     3.546 5.573e+00     0.014
5.546     0.883
## AGE                      -0.001     0.001    -0.722 4.320e+00     0.507
25.461     0.973
## SEXMale                   0.009     0.006     1.677 8.841e+02     0.094
0.072     0.069
## SEXUnknown               -0.199     0.060    -3.320 4.136e+04     0.001
0.010     0.010
## RACEBlack                -0.011     0.022    -0.498 1.723e+03     0.618
0.051     0.049
## RACEOther                -0.031     0.024    -1.304 2.130e+05     0.192
0.004     0.004
## RACEUnknown              -0.028     0.031    -0.910 1.646e+04     0.363
0.016     0.016
## RACEWhite                -0.014     0.016    -0.855 2.540e+03     0.392
0.041     0.040
## TUMOR_PURITY              0.003     0.001     5.055 4.258e+00     0.006
31.547     0.978
## SMOKING_HISTORYFormer    -0.020     0.011    -1.764 1.821e+03     0.078
0.049     0.048
## SMOKING_HISTORYNever     -0.036     0.013    -2.894 2.681e+02     0.004
0.139     0.129
##
## Unadjusted hypothesis test as appropriate in larger samples.
```

From the integrated fixed effect we can find:

- • The estimate of Age is -0.001 with a p-value of 0.507, indicating no significant linear relationship between age and TMB. While the negative coefficient suggests a slight decrease in TMB with age, this effect is not statistically significant.

• The estimate for males is 0.009, with a p-value of 0.094. Being male slightly increases FGA compared to females, but this effect is not statistically significant ($p > 0.05$). • The estimate for unknown sex is -0.199, with a p-value of 0.001. Individuals with unknown sex have significantly lower FGA compared to females, and this effect is statistically significant ($p < 0.05$).

• For the different race categories (Black, Other, Unknown, White), the estimates and p-values indicate that race does not have a significant effect on FGA. All p-values are greater than 0.05, suggesting that race differences are not statistically significant in this model.

• The estimate for TUMOR_PURITY is 0.003 with a p-value of 0.006, indicating a significant positive relationship between tumor purity and FGA. As tumor purity increases, FGA significantly increases.

• The estimate for "Former smoker" is -0.020, with a p-value of 0.078. Former smokers tend to have slightly lower FGA compared to current smokers, but this effect is not statistically significant ($p > 0.05$). • The estimate for "Never smoker" is -0.036, with a p-value of 0.004.Individuals who have never smoked have significantly lower FGA compared to current smokers ($p < 0.05$).

```
# Extracting random effect
## Extraction variance
variance_list <- list()

for (i in 1:5) {
  model_name <- get(paste0("mixed_model_FGA_x_", i))

  re_variance <- as.data.frame(VarCorr(model_name))

  re_variance_df <- data.frame(
    group = re_variance$grp,
    term = re_variance$var1,
    variance = re_variance$vcov
  )

  variance_list[[i]] <- rbind(re_variance_df)
}

variance_df <- do.call(rbind, variance_list)

# Variance of each variables (按 group 和 term 分组)
mean_variances <- aggregate(variance ~ group , data = variance_df, FUN =
mean)

print(mean_variances)
```

```
##         group    variance
## 1 CANCER_TYPE 0.004632224
```

```
## 2  PATIENT_ID 0.029711297
## 3    Residual 0.012134673
## 4       Study 0.005462142
```

From the integrated random effect, we can find:

- Variance of Cancer Type is 0.0046, which attributed to differences between cancer types is relatively small, indicating that there are modest differences in FGA values between various cancer types.

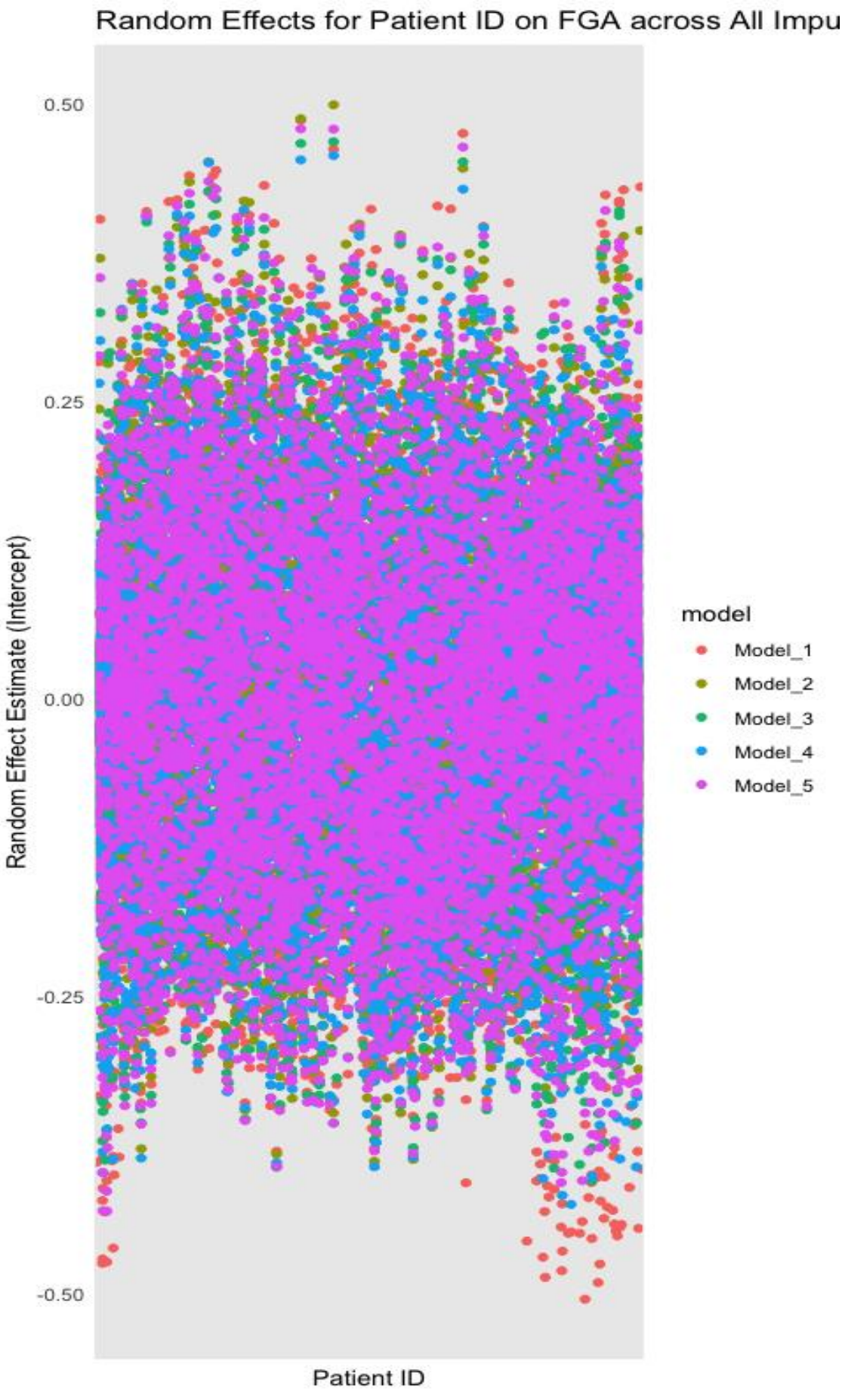
- Variance of Patient ID is 0.0297. The patient-level variance is larger compared to cancer type, suggesting that individual patient characteristics explain more of the variation in FGA compared to cancer type.

- Variance of Study is 0.0055. The variance attributed to the different studies is also relatively small, suggesting that study-specific differences contribute modestly to the variation in FGA.

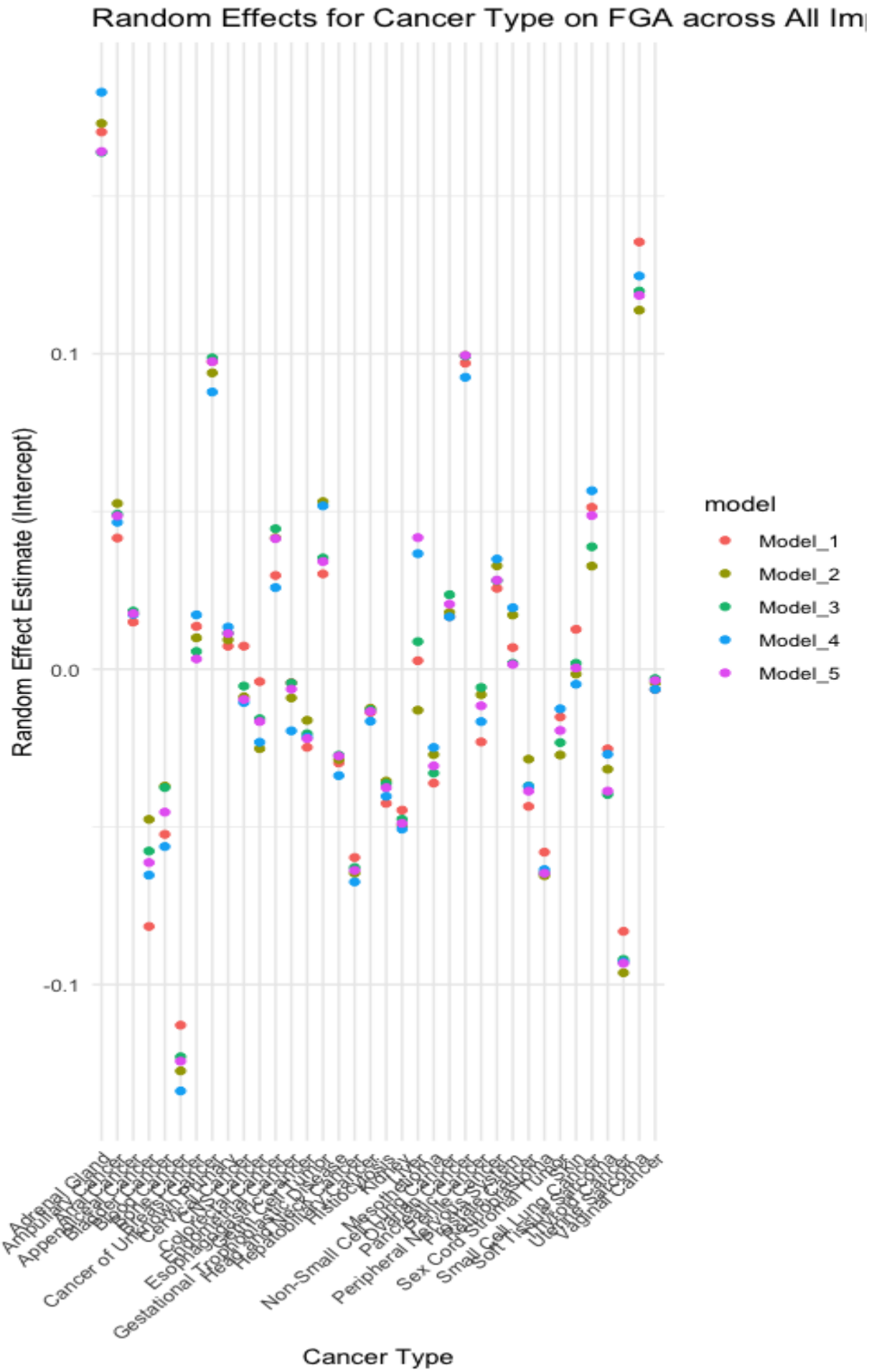
- Variance of residual is 0.0121. The residual variance represents the variation that is not explained by the random effects or the fixed effects in the model. This indicates a moderate level of unexplained variation in FGA.

```
## Extraction intercept (visualizaion all datasets' randam effect on a single plot)
### Extract the random effects of each model and add the model label
extract_random_effects <- function(model, model_name) {
  random_effects <- ranef(model)
  #  get Patient_ID, Cancer_Type, Study random effect
  patient_id_effects <- data.frame(PatientID = rownames(random_effects$PATIENT_ID),
                                   Effect = random_effects$PATIENT_ID$`(Intercept)`,
                                   group = "Patient_ID",
                                   model = model_name)

  cancer_type_effects <- data.frame(CancerType = rownames(random_effects$CANCER_TYPE),
                                    Effect = random_effects$CANCER_TYPE$`(Intercept)`,
                                    group = "Cancer_Type",
                                    model = model_name)

  study_effects <- data.frame(Study = rownames(random_effects$Study),
                              Effect = random_effects$Study$`(Intercept)`,
                              group = "Study",
                              model = model_name)

  return(list(patient_id = patient_id_effects, cancer_type = cancer_type_effects, study = study_effects))
}
```

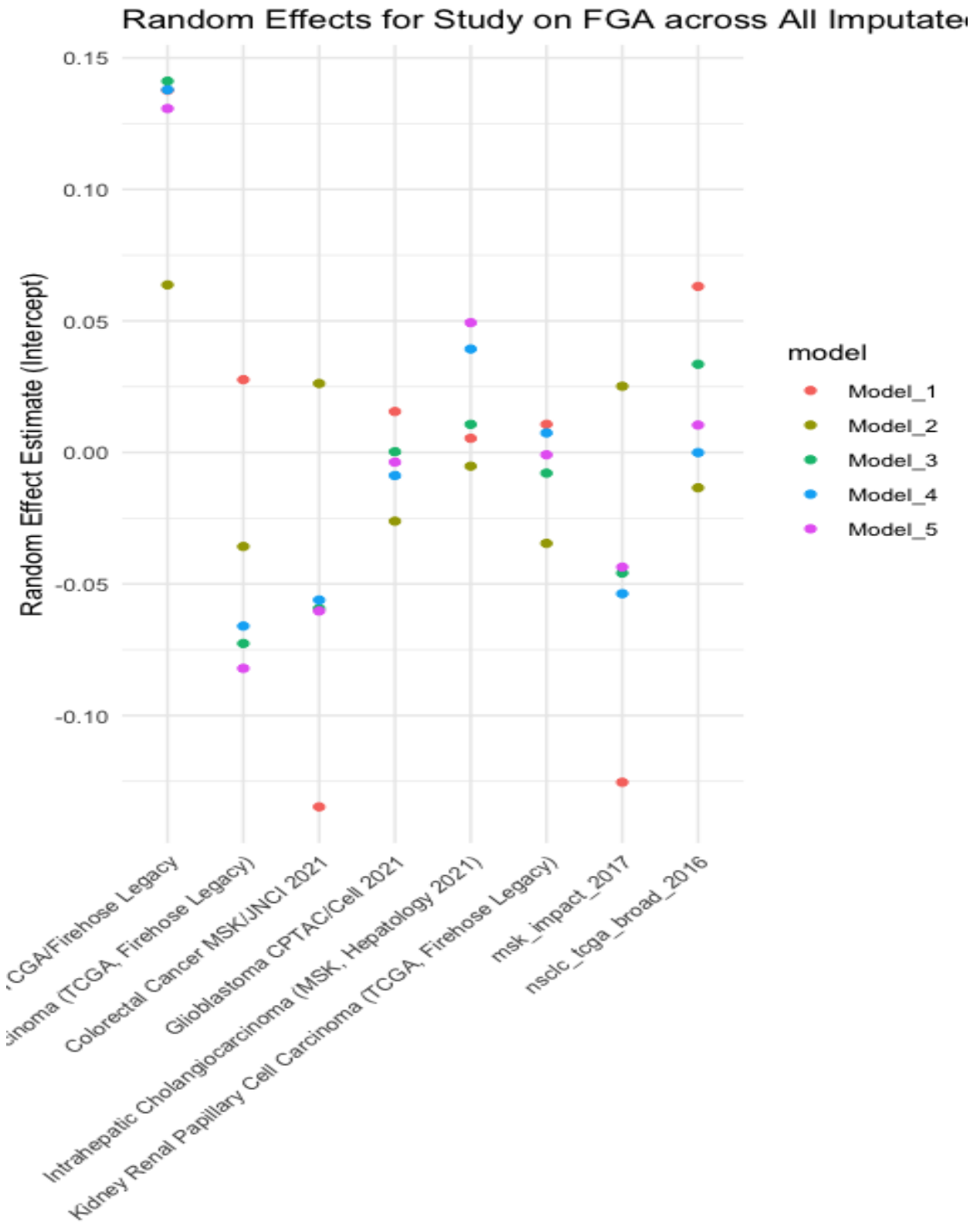
```r
# get 5 models‘ random eff
random_effects_1 <- extract_random_effects(mixed_model_FGA_x_1, "Model_1")
random_effects_2 <- extract_random_effects(mixed_model_FGA_x_2, "Model_2")
random_effects_3 <- extract_random_effects(mixed_model_FGA_x_3, "Model_3")
random_effects_4 <- extract_random_effects(mixed_model_FGA_x_4, "Model_4")
random_effects_5 <- extract_random_effects(mixed_model_FGA_x_5, "Model_5")

# put togehter by vairbale name
combined_patient_id <- rbind(random_effects_1$patient_id,
random_effects_2$patient_id, random_effects_3$patient_id,
                             random_effects_4$patient_id,
random_effects_5$patient_id)

combined_cancer_type <- rbind(random_effects_1$cancer_type,
random_effects_2$cancer_type, random_effects_3$cancer_type,
                              random_effects_4$cancer_type,
random_effects_5$cancer_type)

combined_study <- rbind(random_effects_1$study, random_effects_2$study,
random_effects_3$study,
                        random_effects_4$study, random_effects_5$study)

# Visual。。。。
ggplot(combined_patient_id, aes(x = PatientID, y = Effect, color = model)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Patient ID on FGA across All Imputated
Datasets",
       x = "Patient ID", y = "Random Effect Estimate (Intercept)") +
  # theme(axis.text.x = element_text(angle = 45, hjust = 1))
  theme(axis.text.x = element_blank())
```


Random Effects for Patient ID on FGA across All Impu
0.50
0.25
0.00
-0.25
-0.50
Random Effect Estimate (Intercept)
Patient ID
model
Model_1
Model_2
Model_3
Model_4
Model_5

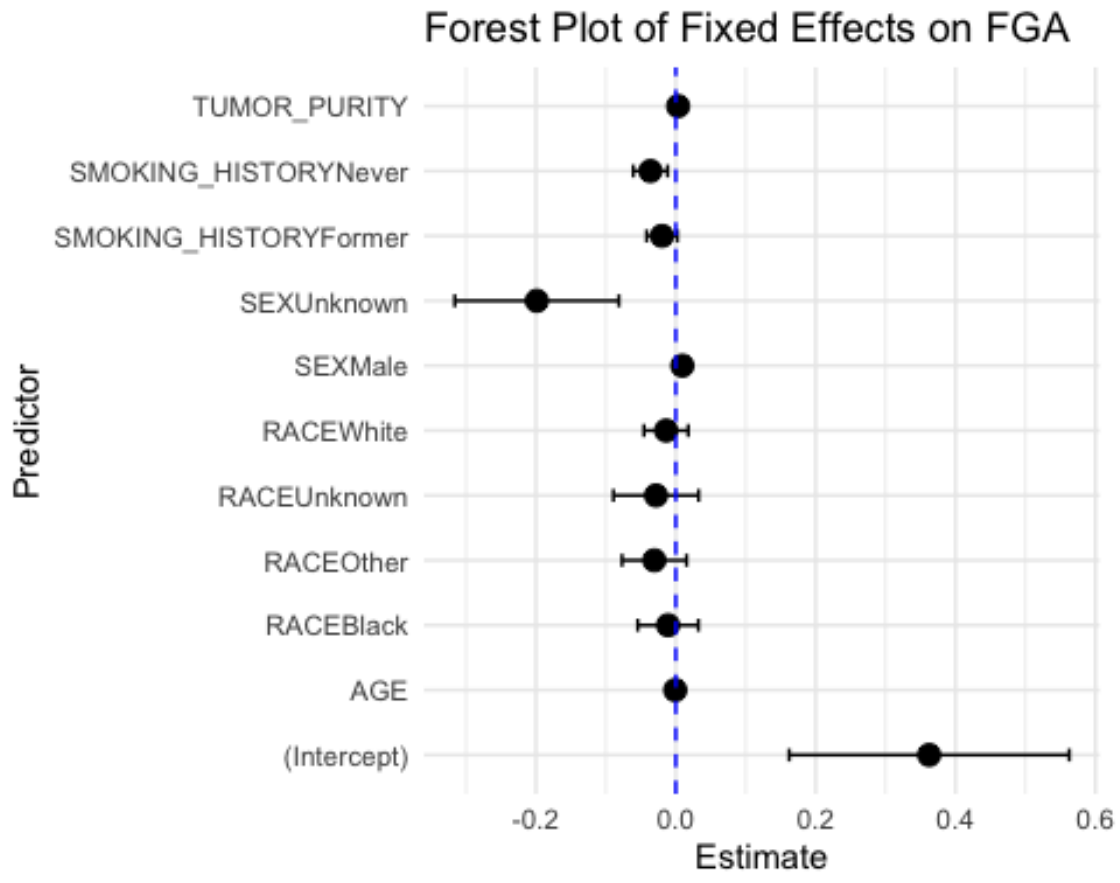
```r
ggplot(combined_cancer_type, aes(x = CancerType, y = Effect, color = model))
+
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Cancer Type on FGA across All Imputated
Datasets",
       x = "Cancer Type", y = "Random Effect Estimate (Intercept)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Random Effects for Cancer Type on FGA across All Im
Random Effect Estimate (Intercept)
0.1
0.0
-0.1
Cancer Type
model
Model_1
Model_2
Model_3
Model_4
Model_5

```r
ggplot(combined_study, aes(x = Study, y = Effect, color = model)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Random Effects for Study on FGA across All Imputated
Datasets",
       x = "Study", y = "Random Effect Estimate (Intercept)") +
  scale_y_continuous(limits = c(min(combined_study$Effect),
max(combined_study$Effect))) +
  # theme(axis.text.x = element_blank())
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Random Effects for Study on FGA across All Imputate
Random Effect Estimate (Intercept)
0.15
0.10
0.05
0.00
-0.05
-0.10
model
Model_1
Model_2
Model_3
Model_4
Model_5
CGA/Firehose Legacy
inoma (TCGA, Firehose Legacy)
Colorectal Cancer MSK/JNCI 2021
Glioblastoma CPTAC/Cell 2021
Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021)
Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy)
msk_impact_2017
nsclc_tcga_broad_2016
Study

```r
estimates <- pooled_results_FGA_x$estimates

forest_data <- data.frame(
  term = rownames(estimates),
  estimate = estimates[, "Estimate"],
  std.error = estimates[, "Std.Error"],
  conf.low = estimates[, "Estimate"] - 1.96 * estimates[, "Std.Error"],
  conf.high = estimates[, "Estimate"] + 1.96 * estimates[, "Std.Error"]
)


ggplot(forest_data, aes(x = estimate, y = term)) +
  geom_point(size = 3) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  theme_minimal() +
  labs(title = "Forest Plot of Fixed Effects on FGA", x = "Estimate", y =
"Predictor") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "blue")
```



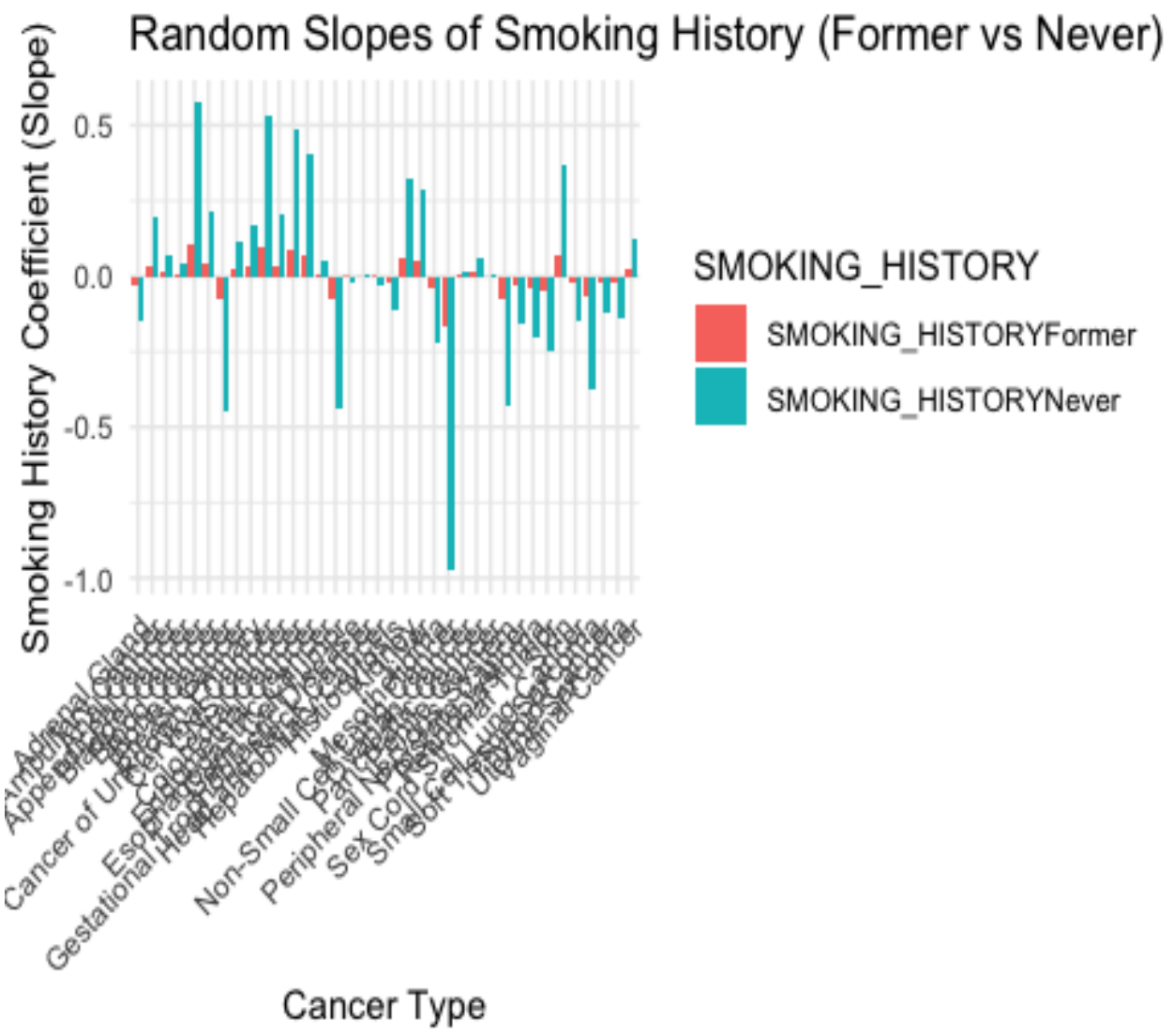
Smoking History vs Cancer Type

Our Additional Model with Random Slope of Smoking Effect on Cancer Type is：

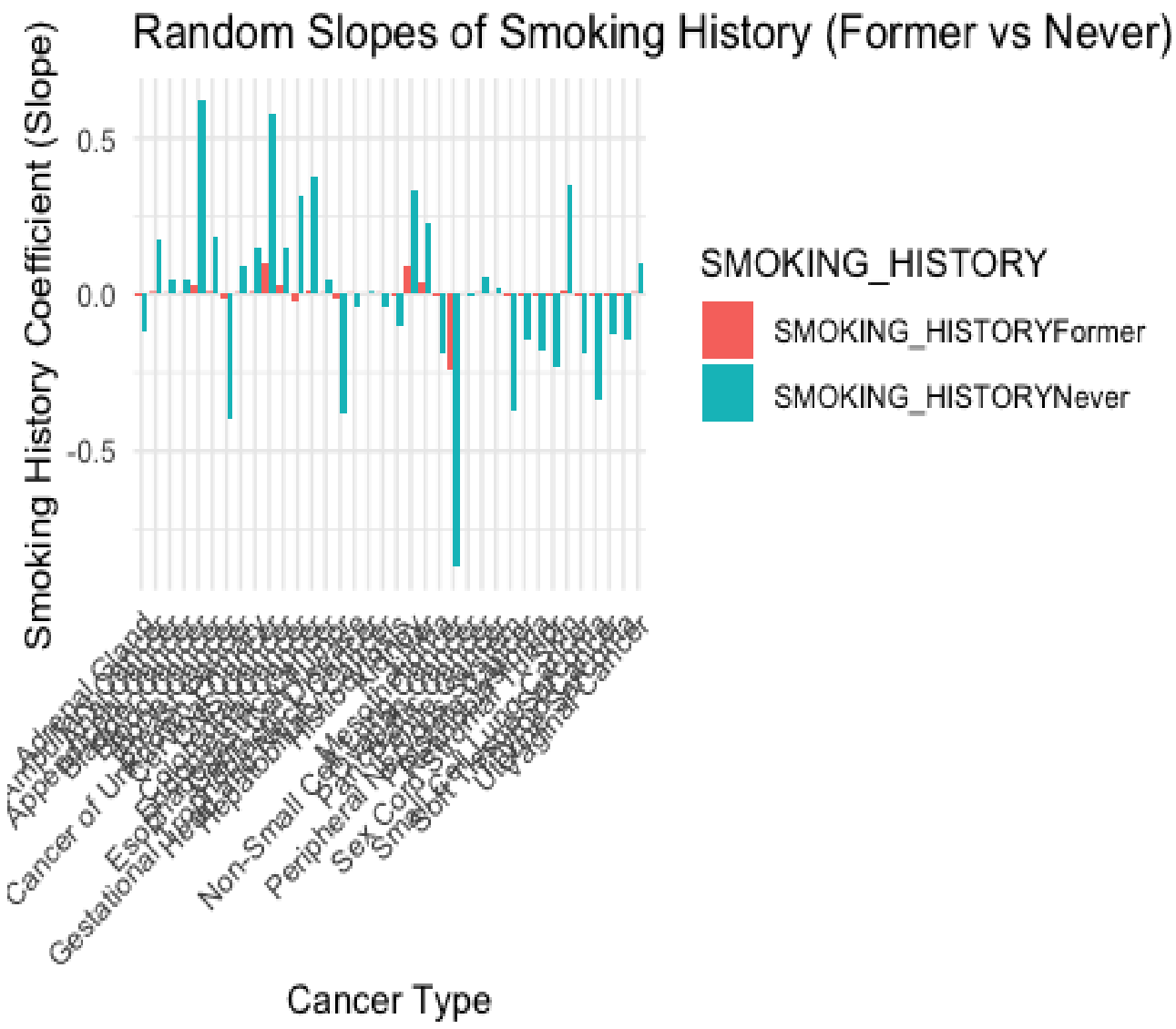
$$\log_2(\text{TMB}+1) = \beta_0 + \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{SEX} + \beta_3 \cdot \text{RACE} + \beta_4 \cdot \text{TUMOR\_PURITY} + \beta_5 \cdot \$\$ + u_{\text{PATIENT\_ID}} + u_{\text{Study}} + u_{\text{CANCER\_TYPE}} + u_{\text{CANCER\_TYPE}} \cdot \text{SMOKING\_HISTORY} + \epsilon$$

```r
# VIsualiaztion for smoking history in different cancer type
# Extract the random effects
random_effects_1 <- ranef(SH_CT_TMB_1)$CANCER_TYPE
random_effects_2 <- ranef(SH_CT_TMB_2)$CANCER_TYPE
random_effects_3 <- ranef(SH_CT_TMB_3)$CANCER_TYPE
random_effects_4 <- ranef(SH_CT_TMB_4)$CANCER_TYPE
random_effects_5 <- ranef(SH_CT_TMB_5)$CANCER_TYPE

# Extract the random slopes of SH
smoking_slope_1 <- as.data.frame(random_effects_1)[,
c("SMOKING_HISTORYFormer", "SMOKING_HISTORYNever")]
smoking_slope_2 <- as.data.frame(random_effects_2)[,
c("SMOKING_HISTORYFormer", "SMOKING_HISTORYNever")]
smoking_slope_3 <- as.data.frame(random_effects_3)[,
c("SMOKING_HISTORYFormer", "SMOKING_HISTORYNever")]
smoking_slope_4 <- as.data.frame(random_effects_4)[,
c("SMOKING_HISTORYFormer", "SMOKING_HISTORYNever")]
smoking_slope_5 <- as.data.frame(random_effects_5)[,
c("SMOKING_HISTORYFormer", "SMOKING_HISTORYNever")]


smoking_slope_1$CANCER_TYPE <- rownames(random_effects_1)
smoking_slope_2$CANCER_TYPE <- rownames(random_effects_2)
smoking_slope_3$CANCER_TYPE <- rownames(random_effects_3)
smoking_slope_4$CANCER_TYPE <- rownames(random_effects_4)
smoking_slope_5$CANCER_TYPE <- rownames(random_effects_5)

# print(smoking_slope)



# to long format
smoking_slope_long_1 <- gather(smoking_slope_1, key = "SMOKING_HISTORY",
value = "Slope", -CANCER_TYPE)
smoking_slope_long_2 <- gather(smoking_slope_2, key = "SMOKING_HISTORY",
value = "Slope", -CANCER_TYPE)
smoking_slope_long_3 <- gather(smoking_slope_3, key = "SMOKING_HISTORY",
value = "Slope", -CANCER_TYPE)
smoking_slope_long_4 <- gather(smoking_slope_4, key = "SMOKING_HISTORY",
value = "Slope", -CANCER_TYPE)
smoking_slope_long_5 <- gather(smoking_slope_5, key = "SMOKING_HISTORY",
value = "Slope", -CANCER_TYPE)

# print(smoking_slope_long)
```

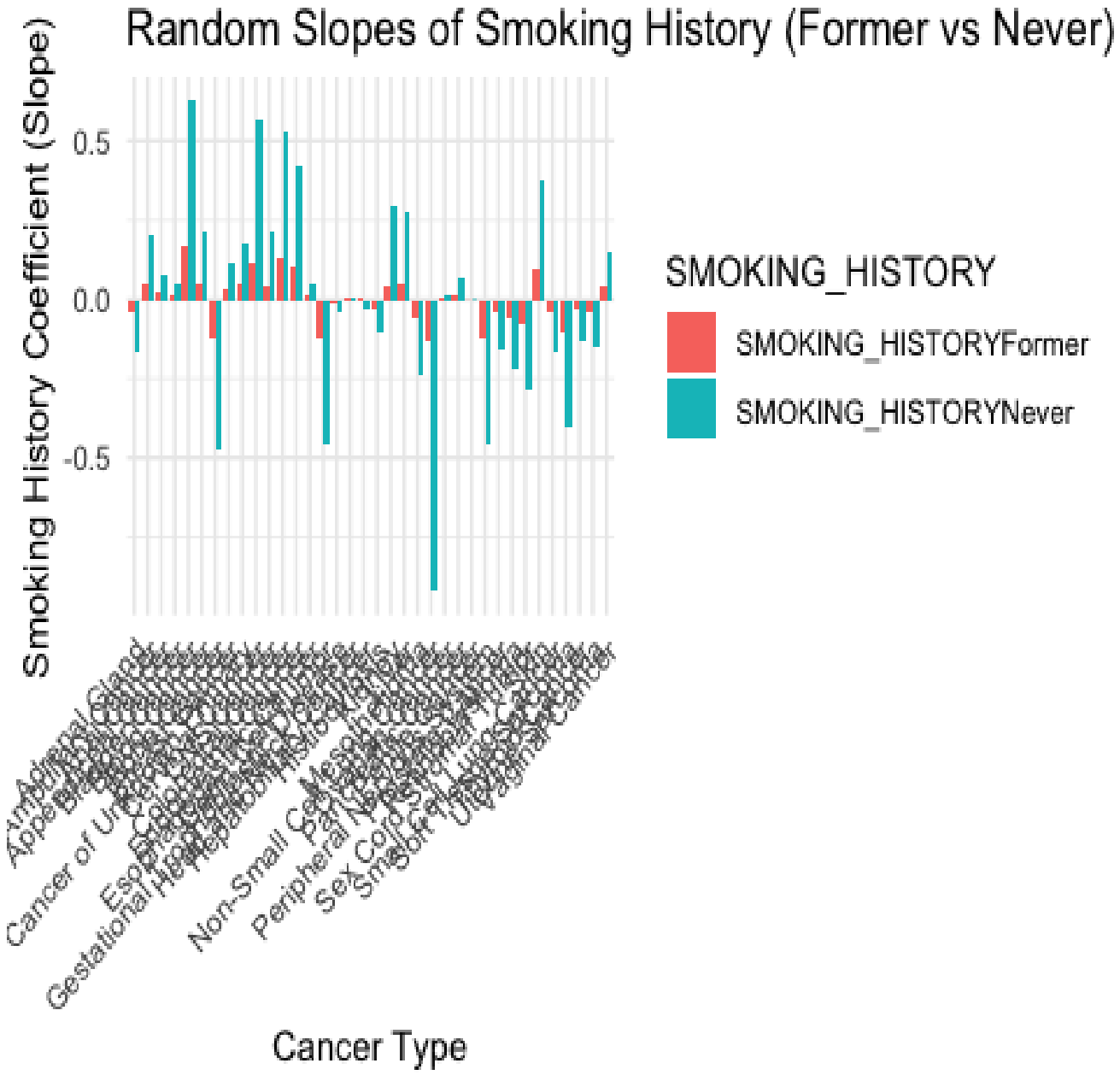
```r
ggplot(smoking_slope_long_1, aes(x = CANCER_TYPE, y = Slope, fill =
SMOKING_HISTORY)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Cancer Type", y = "Smoking History Coefficient (Slope)",
       title = "Random Slopes of Smoking History (Former vs Never) across
Cancer Types plt#1") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



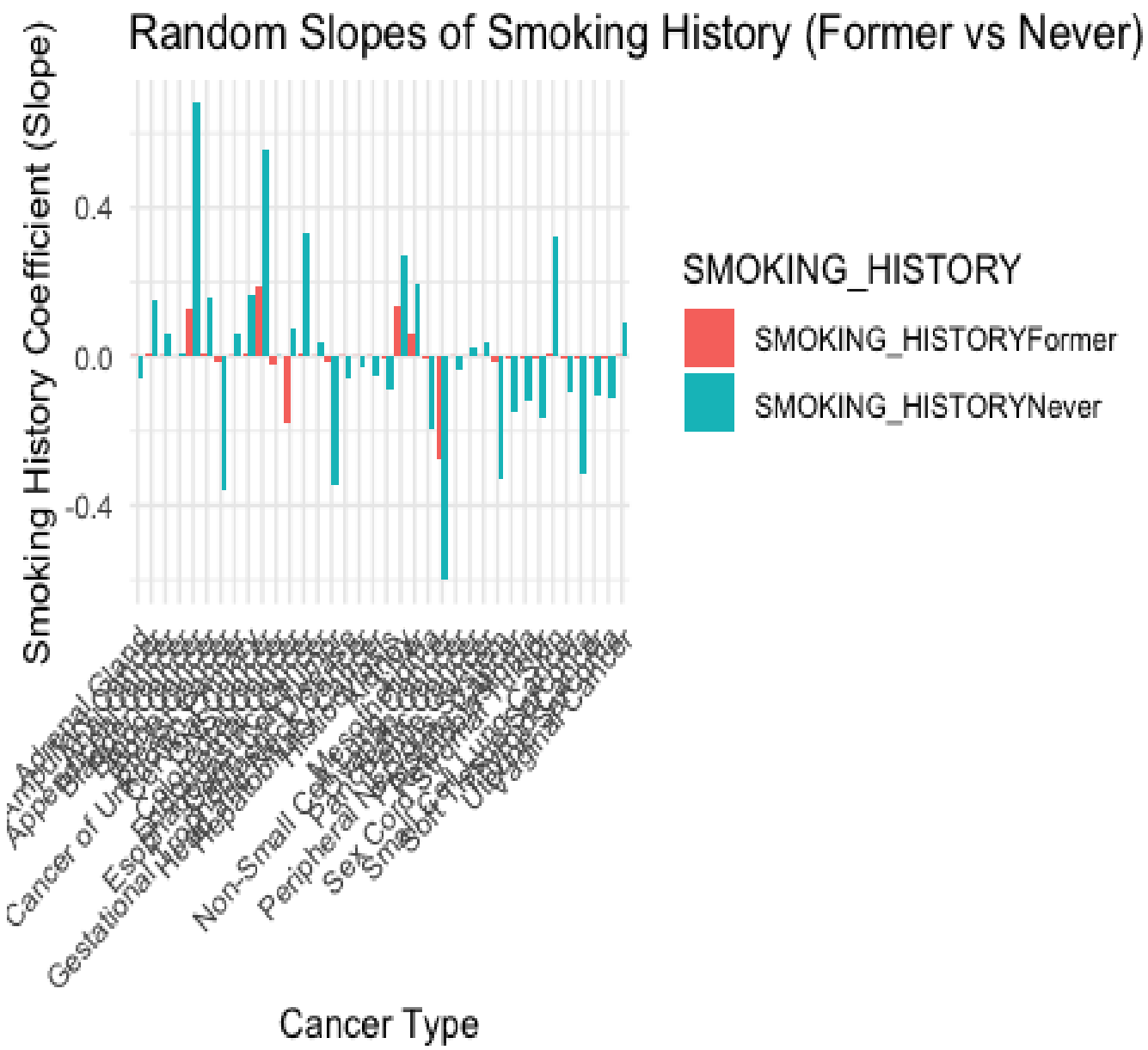

```r
ggplot(smoking_slope_long_2, aes(x = CANCER_TYPE, y = Slope, fill =
SMOKING_HISTORY)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Cancer Type", y = "Smoking History Coefficient (Slope)",
       title = "Random Slopes of Smoking History (Former vs Never) across
Cancer Types plt#2") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



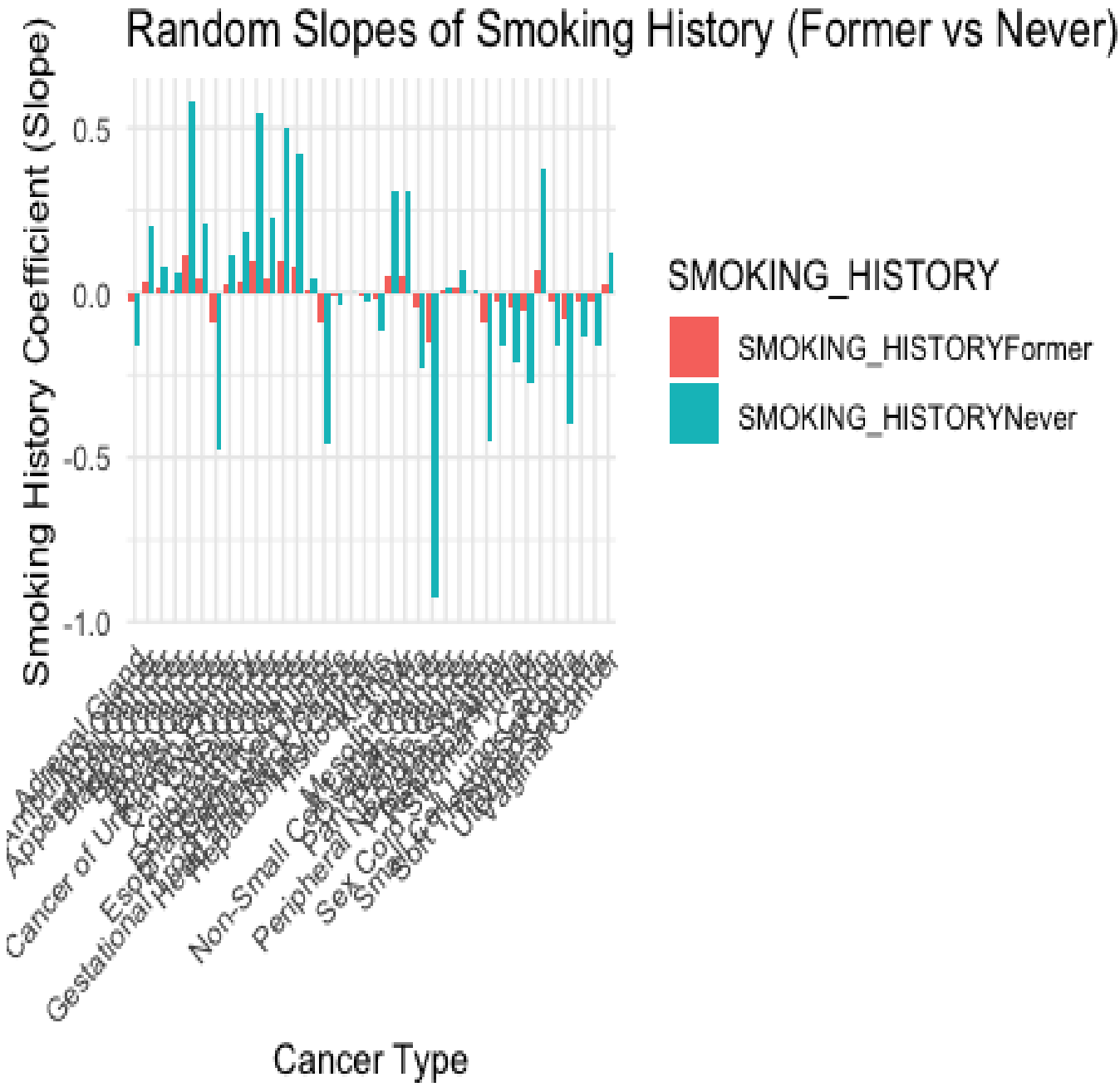

```r
ggplot(smoking_slope_long_3, aes(x = CANCER_TYPE, y = Slope, fill =
SMOKING_HISTORY)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Cancer Type", y = "Smoking History Coefficient (Slope)",
       title = "Random Slopes of Smoking History (Former vs Never) across
Cancer Types plt#3") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
ggplot(smoking_slope_long_4, aes(x = CANCER_TYPE, y = Slope, fill =
SMOKING_HISTORY)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Cancer Type", y = "Smoking History Coefficient (Slope)",
       title = "Random Slopes of Smoking History (Former vs Never) across
Cancer Types plt#4") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

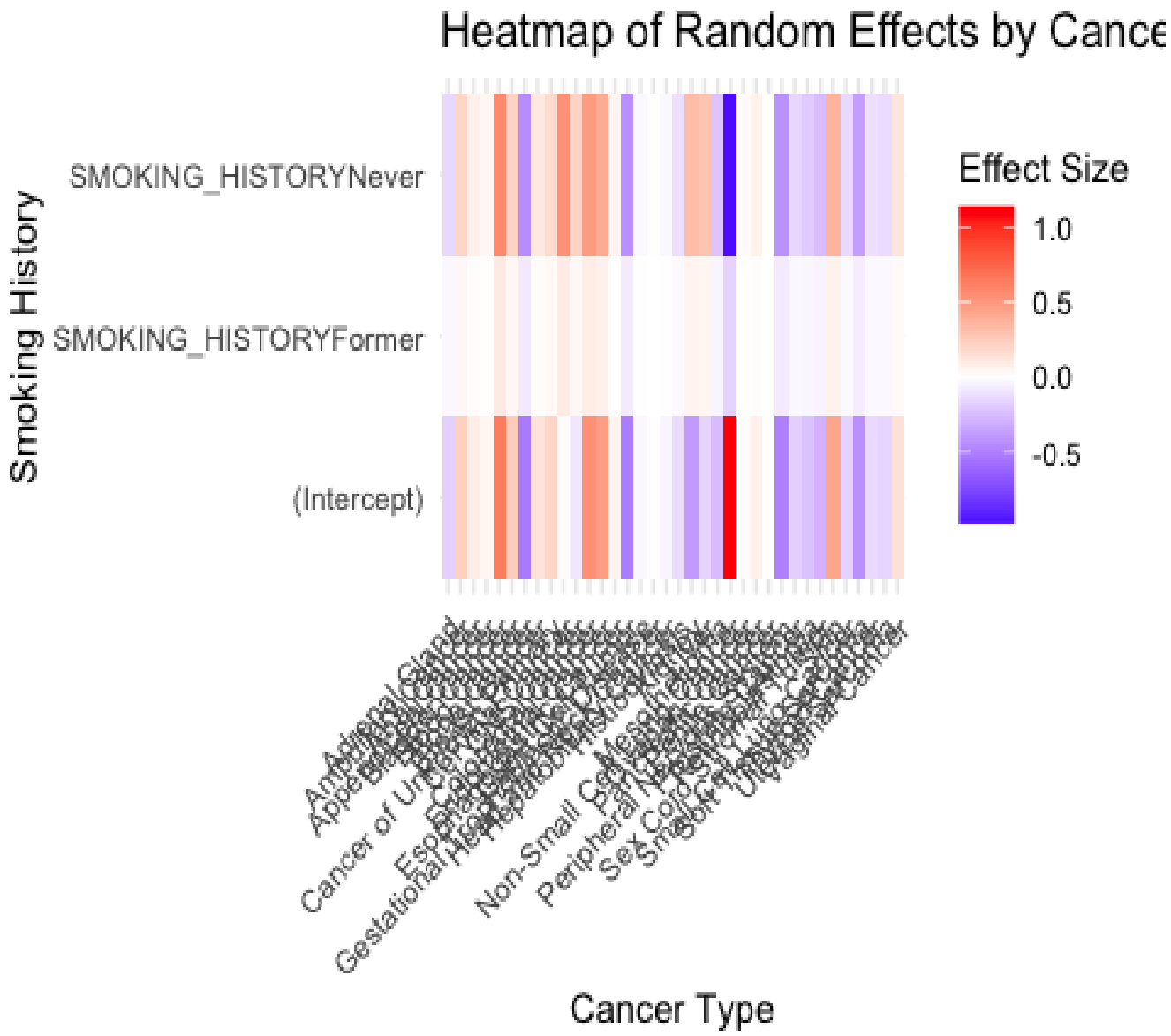



```
ggplot(smoking_slope_long_5, aes(x = CANCER_TYPE, y = Slope, fill =
SMOKING_HISTORY)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Cancer Type", y = "Smoking History Coefficient (Slope)",
       title = "Random Slopes of Smoking History (Former vs Never) across
Cancer Types plt#5") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

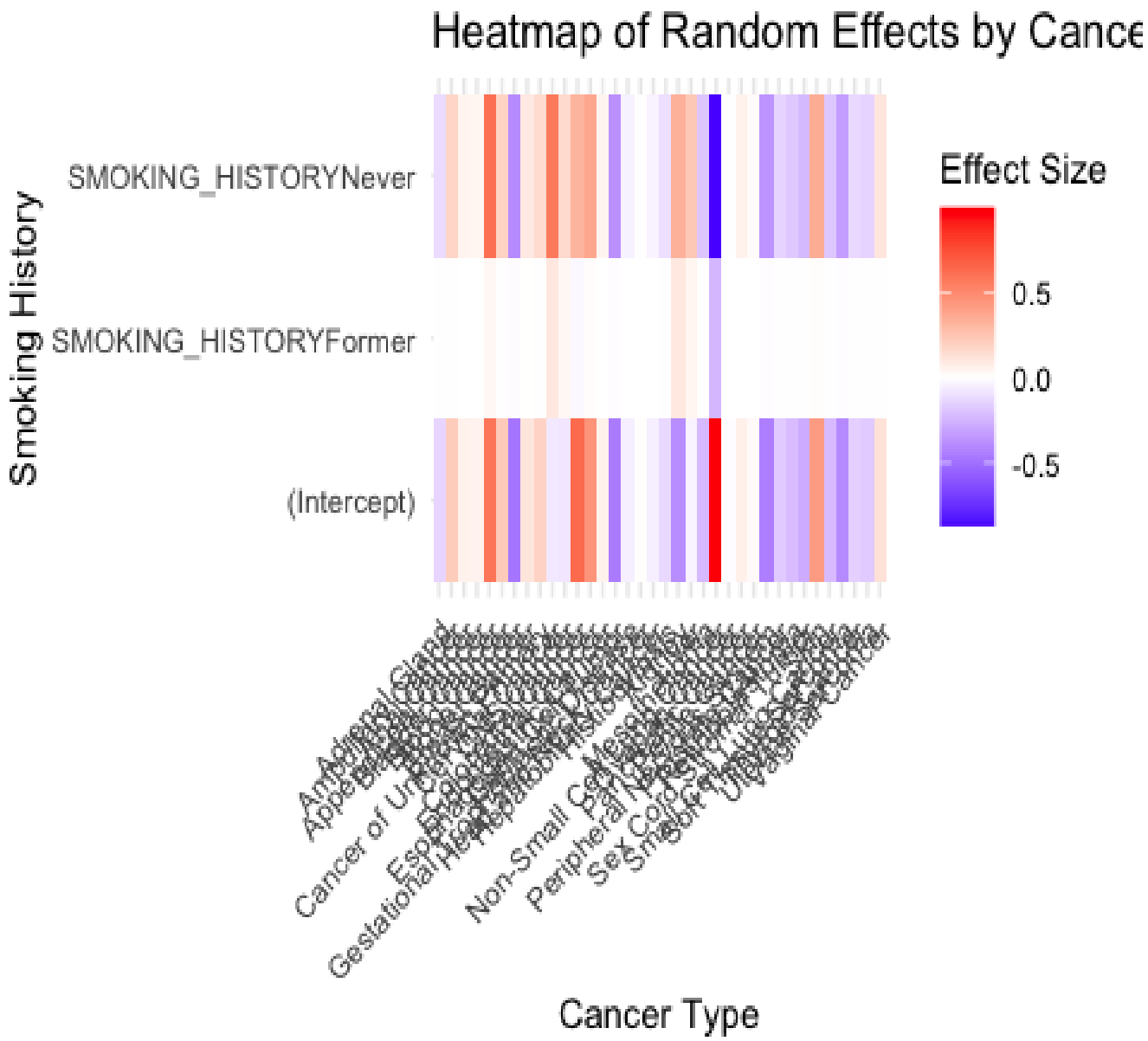



```r
## Heat map
ranef_effects_1 <- ranef(SH_CT_TMB_1)
ranef_effects_2 <- ranef(SH_CT_TMB_2)
ranef_effects_3 <- ranef(SH_CT_TMB_3)
ranef_effects_4 <- ranef(SH_CT_TMB_4)
ranef_effects_5 <- ranef(SH_CT_TMB_5)

random_effects_df_1 <- as.data.frame(ranef_effects_1$CANCER_TYPE)
random_effects_df_2 <- as.data.frame(ranef_effects_2$CANCER_TYPE)
random_effects_df_3 <- as.data.frame(ranef_effects_3$CANCER_TYPE)
```

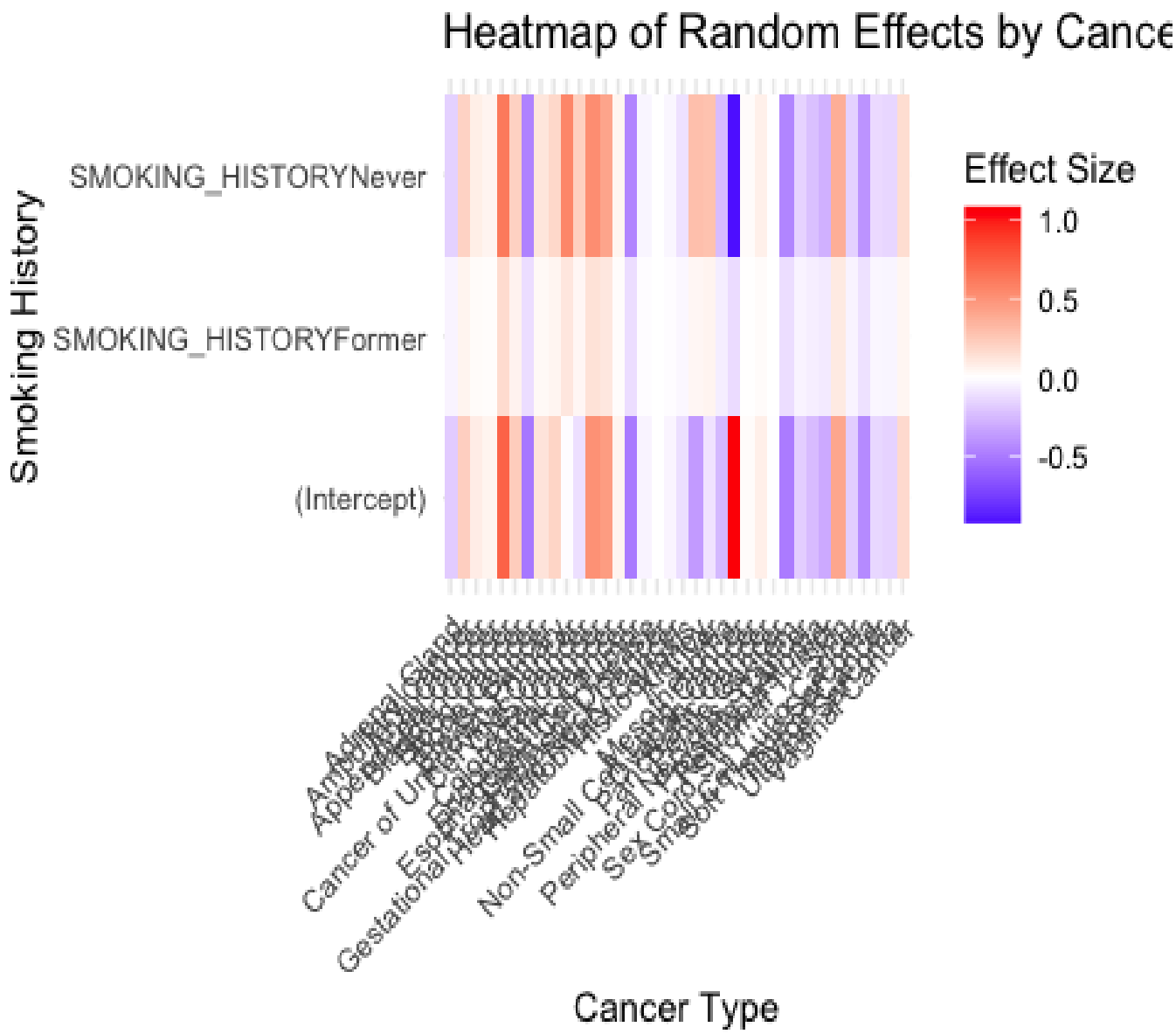
```r
random_effects_df_4 <- as.data.frame(ranef_effects_4$CANCER_TYPE)
random_effects_df_5 <- as.data.frame(ranef_effects_5$CANCER_TYPE)

random_effects_df_1$cancer_type <- rownames(random_effects_df_1)
random_effects_df_2$cancer_type <- rownames(random_effects_df_2)
random_effects_df_3$cancer_type <- rownames(random_effects_df_3)
random_effects_df_4$cancer_type <- rownames(random_effects_df_4)
random_effects_df_5$cancer_type <- rownames(random_effects_df_5)


random_effects_df_1 <- random_effects_df_1 %>%
  select(cancer_type, `(Intercept)`, SMOKING_HISTORYFormer,
SMOKING_HISTORYNever)
random_effects_df_2 <- random_effects_df_2 %>%
  select(cancer_type, `(Intercept)`, SMOKING_HISTORYFormer,
SMOKING_HISTORYNever)
random_effects_df_3 <- random_effects_df_3 %>%
  select(cancer_type, `(Intercept)`, SMOKING_HISTORYFormer,
SMOKING_HISTORYNever)
random_effects_df_4 <- random_effects_df_4 %>%
  select(cancer_type, `(Intercept)`, SMOKING_HISTORYFormer,
SMOKING_HISTORYNever)
random_effects_df_5 <- random_effects_df_5 %>%
  select(cancer_type, `(Intercept)`, SMOKING_HISTORYFormer,
SMOKING_HISTORYNever)

random_effects_long_1 <- pivot_longer(random_effects_df_1, cols = -
cancer_type, names_to = "Smoking_History", values_to = "Effect_Size")
random_effects_long_2 <- pivot_longer(random_effects_df_2, cols = -
cancer_type, names_to = "Smoking_History", values_to = "Effect_Size")
random_effects_long_3 <- pivot_longer(random_effects_df_3, cols = -
cancer_type, names_to = "Smoking_History", values_to = "Effect_Size")
random_effects_long_4 <- pivot_longer(random_effects_df_4, cols = -
cancer_type, names_to = "Smoking_History", values_to = "Effect_Size")
random_effects_long_5 <- pivot_longer(random_effects_df_5, cols = -
cancer_type, names_to = "Smoking_History", values_to = "Effect_Size")

ggplot(random_effects_long_1, aes(x = cancer_type, y = Smoking_History, fill
= Effect_Size)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
  theme_minimal() +
  labs(title = "Heatmap of Random Effects by Cancer Type and Smoking History
plt#1",
       x = "Cancer Type",
       y = "Smoking History",
       fill = "Effect Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



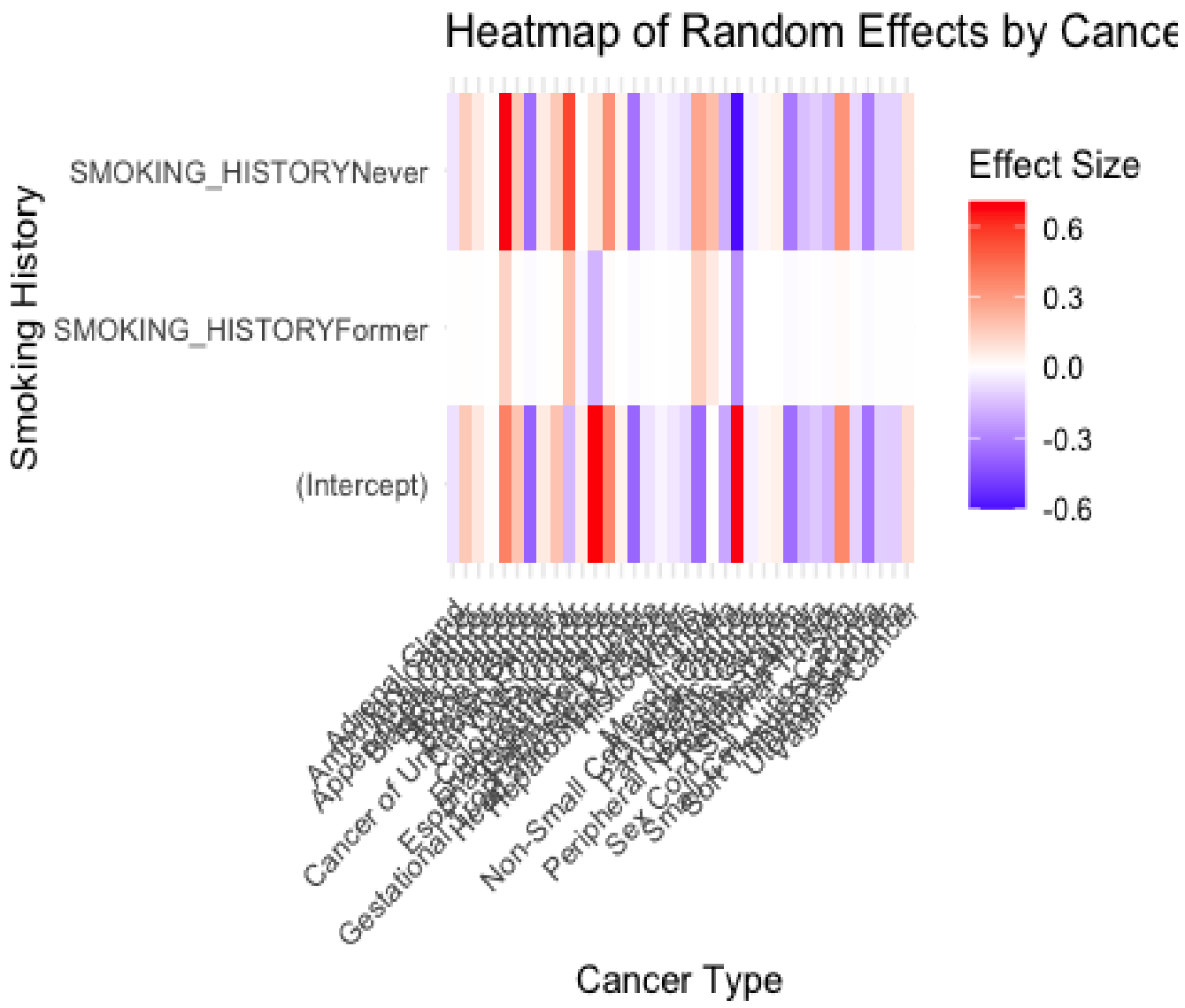

```
ggplot(random_effects_long_2, aes(x = cancer_type, y = Smoking_History, fill
= Effect_Size)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
  theme_minimal() +
  labs(title = "Heatmap of Random Effects by Cancer Type and Smoking History
plt#2",
       x = "Cancer Type",
       y = "Smoking History",
       fill = "Effect Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



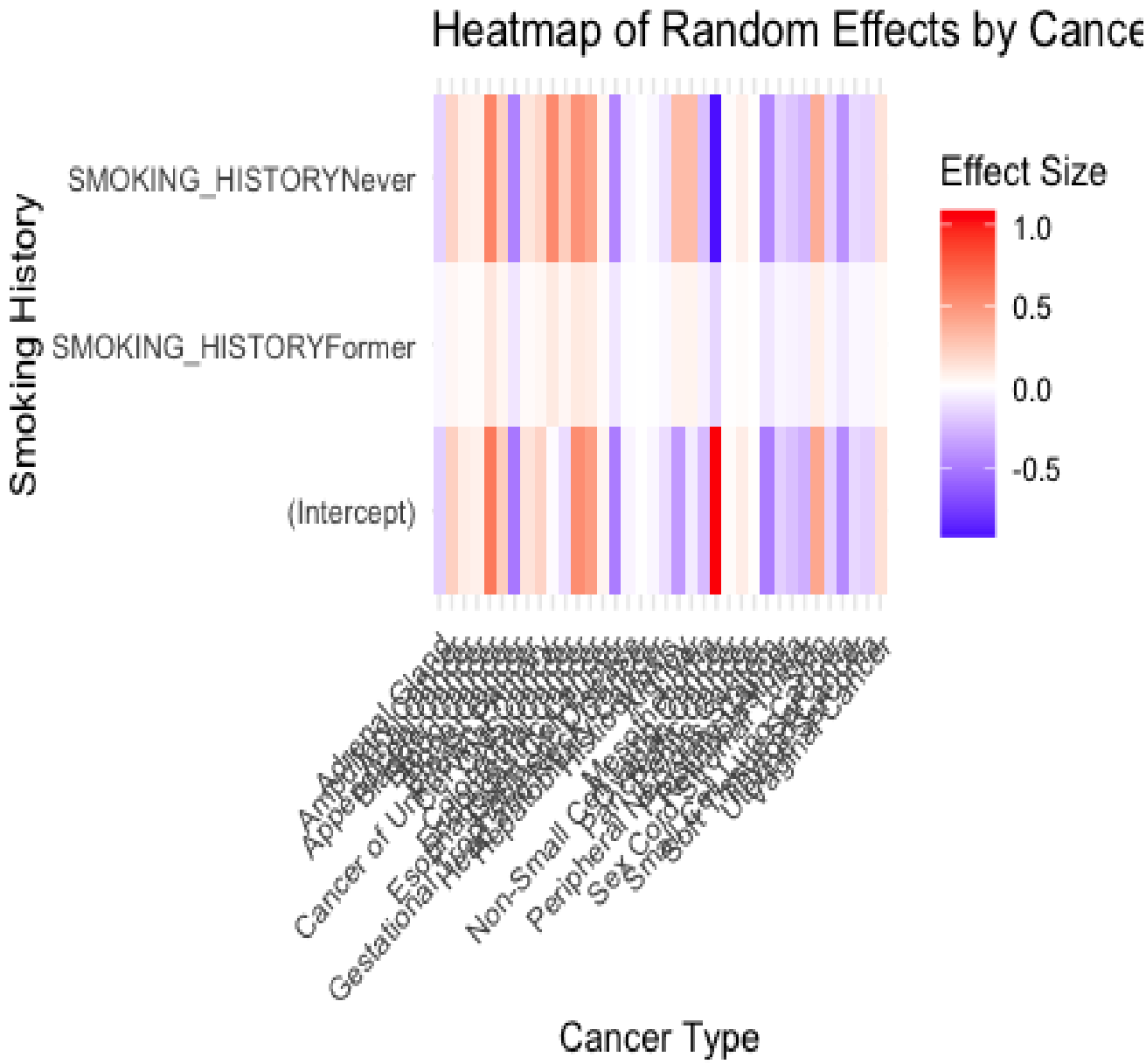

```r
ggplot(random_effects_long_3, aes(x = cancer_type, y = Smoking_History, fill = Effect_Size)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Heatmap of Random Effects by Cancer Type and Smoking History plt#3",
       x = "Cancer Type",
       y = "Smoking History",
       fill = "Effect Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(random_effects_long_4, aes(x = cancer_type, y = Smoking_History, fill
= Effect_Size)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
  theme_minimal() +
  labs(title = "Heatmap of Random Effects by Cancer Type and Smoking History
plt#4",
       x = "Cancer Type",
       y = "Smoking History",
       fill = "Effect Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(random_effects_long_5, aes(x = cancer_type, y = Smoking_History, fill
= Effect_Size)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
  theme_minimal() +
  labs(title = "Heatmap of Random Effects by Cancer Type and Smoking History
plt#5",
       x = "Cancer Type",
       y = "Smoking History",
       fill = "Effect Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the Forest plots over all the imputed data sets, we can see that the effect of Smoking History varies across different cancer types. The effect can be significantly larger on some specific cancer type.

From the heat maps over all the imputed data sets, we can see the effect of Smoking History on Non-Small Cell Lung Cancer is specifically significant compared with the one on other cancer. Compared to current smokers, the former smokers and people who never smoked have a significant negative effect on transformed TMB.

Discussion & Conclusion

This study investigated the effects of demographic, clinical, and genomic factors on two important indicators of tumor genome alterations: Tumor Mutation Burden (TMB) and Fraction of Genome Altered (FGA). Through the application of mixed-effects models across five imputed datasets, both fixed and random effects were analyzed to assess the variability at the patient, cancer type, and study levels. The models provided insights into how age, sex, race, tumor purity, and smoking history influence TMB and FGA. We also compared the differences in model fit and interpretability across various hierarchical structures of random effects. Special attention was given to the interaction between smoking history and cancer type, analyzing how this interaction influences variations in TMB.

Despite testing various model combinations, including nested and crossed random effects, we ultimately found that models retaining the crossed structure for patient ID, study, and cancer type as random effects performed best. This model demonstrated better fit according to AIC and BIC metrics, with a conditional R-squared of 0.853 (for TMB) and 0.793 (for FGA), indicating that it effectively explains the variability in genomic alterations.

It is important to note that during the model diagnostics, the residuals for both the FGA and TMB models did not follow a perfectly normal distribution. The Q-Q plots showed some degree of deviation at the tails, indicating that our models still exhibit a certain level of bias. Therefore, in future research, we will explore additional modeling approaches to achieve better fit, including models based on alternative distributional assumptions or Bayesian models.

For TMB, the fixed-effects analysis revealed that sex has a significant effect, indicating that male patients tend to have higher TMB values than females. Smoking history was also found to significantly influence TMB, with former smokers showing lower TMB compared to current-smokers and the never-smoker showing the lowest TMB. Tumor purity had a positive and significant association with TMB, suggesting that purer tumor samples have higher mutation burdens. Notably, race and age did not demonstrate a significant effect on TMB. The random effects showed considerable variance at the patient and cancer type levels, reflecting the heterogeneity of TMB within and across different cancer types.

For FGA, the fixed-effects analysis showed that tumor purity was positively and significantly associated with FGA, consistent with the biological expectation that purer tumors exhibit higher fractions of genome alteration. However, age, sex, and race did not have significant effects on FGA. Smoking history, while not significant for former-smokers compared to current smoker, but the never smokers showed a significant lower TMB. In terms of random effects, the variance was predominantly observed at the patient level, while cancer type and study contributed only modestly to the variation in FGA.

The random effects in both models highlight substantial patient-level variability for TMB and FGA, underscoring the importance of individualized approaches in cancer genomic studies. Differences across cancer types were more pronounced for TMB than for FGA, suggesting that the biological processes driving mutation burden are more cancer-type specific, while genome alteration may be driven by more patient-specific factors.

Additionally, further analysis revealed that the effect of smoking history on transformed TMB varies significantly across different cancer types, with distinct directions and magnitudes in certain cancers. In particular, for Non-Small Cell Lung Cancer (NSCLC), smoking history showed a highly significant effect. Current smoking had a strong positive influence on TMB, while never smoking exhibited a strong negative effect. This highlights the heterogeneity in the relationship between smoking history and TMB across cancer types, with NSCLC demonstrating especially unique patterns.

In summary, TMB is largely influenced by patient sex, tumor purity, and smoking history, with significant variability between patients and cancer types. On the other hand, FGA shows more variability at the patient level and is most strongly associated with tumor purity. These findings suggest that while TMB and FGA are both key genomic indicators, they are influenced by distinct biological and clinical factors, and their variability is shaped differently across patient and cancer type levels. This study revealed the associations between clinical factors and different measurements of genomic alterations. These findings provide deeper insights into how clinical characteristics influence genomic alterations across cancer types, supporting personalized approaches in cancer treatment.

Bibliography

Bladder Urothelial Carcinoma (TCGA, Firehose Legacy). (2020, September 15). Mskcc.org. https://datacatalog.mskcc.org/dataset/10466 Boerner, T., Drill, E., Pak, L. M., Nguyen, B., Sigel, C. S., Alexandre Doussot, Shin, P., Goldman, D. A., Mithat Gönen, Allen, P. J., Balachandran, V. P., Cercek, A., Harding, J. J., Solit, D. B., Schultz, N., Kundra, R., Walch, H., D'Angelica, M. I., DeMatteo, R. P., & Drebin, J. A. (2021). Genetic Determinants of Outcome in Intrahepatic Cholangiocarcinoma. Hepatology, 74(3), 1429–1444. https://doi.org/10.1002/hep.31829 cBioPortal for Cancer Genomics. (n.d.). Www.cbioportal.org. https://www.cbioportal.org/ Cercek, A., Chatila, W. K., Yaeger, R., Walch, H., Dos, G., Krishnan, A., Lerie Palmaira, Maio, A., Kemel, Y., Srinivasan, P., Chaitanya Bandlamudi, Salo-Mullen, E. E., Prince Rainier Tejada, Kimeisha Belanfanti, Galle, J., Vijai, J., Segal, N. H., Varghese, A. M., Diane Lauren Reidy, & Shia, J. (2021). A Comprehensive Comparison of Early-Onset and Average-Onset Colorectal Cancers. 113(12), 1683–1692. https://doi.org/10.1093/jnci/djab124 Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma. (2020, October 20). Mskcc.org. https://datacatalog.mskcc.org/dataset/10476 Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy). (2020, October 27). Mskcc.org. https://datacatalog.mskcc.org/dataset/10480 PanCanAtlas Publications | NCI Genomic

Data Commons. (n.d.). Gdc.cancer.gov. https://gdc.cancer.gov/about-data/publications/pancanatlas Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., & Bajorin, D. F. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nature Genetics, 51(2), 202–206. https://doi.org/10.1038/s41588-018-0312-8 Seldon, C., Karthik Meiyappan, Hoffman, H. I., Guo, J. A., Goel, N., Hwang, W. L., Nguyen, P. L., Mahal, B. A., & Alshalalfa, M. (2022). Genomic alterations predictive of poor clinical outcomes in pan- cancer. Oncotarget, 13(1), 1069–1077. https://doi.org/10.18632/oncotarget.28276 Smith, J., Parl, F. F., & Dupont, W. D. (2023). Mutation Burden Independently Predicts Survival in the Pan-Cancer Atlas. 7. https://doi.org/10.1200/po.22.00571 The Pan-Cancer Atlas. (2013). Cell.com. https://www.cell.com/pb- assets/consortium/PanCancerAtlas/PanCani3/index.html Uterine Corpus Endometrial Carcinoma (TCGA, Firehose Legacy). (2020, November 19). Mskcc.org. https://datacatalog.mskcc.org/dataset/10494 Wang, L.-B., Karpova, A., Gritsenko, M. A., Kyle, J. E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J. H., Hong, R., Stathias, V., Cornwell, M., Petralia, F., Wu, Y., Reva, B., Krug, K., Pugliese, P., Kawaler, E., Olsen, L. K., & Liang, W.-W. (2021). Proteogenomic and metabolomic characterization of human
glioblastoma. Cancer Cell, 39(4), 509-528.e20. https://doi.org/10.1016/j.ccell.2021.01.006 Wan, L., Wang, Z., Xue, J., Yang, H., & Zhu, Y. (2020). Tumor mutation burden predicts response and survival to immune checkpoint inhibitors: a meta-analysis. Translational Cancer Research, 9(9), 5437–5449. https://doi.org/10.21037/tcr-20-1131