**Factors Associated with Genomic Alteration Measurements in Cancerous Tumor Cells**

Amanda Carrico and Yuxiang Ren

Population Health Sciences, Weill Cornell Graduate School of Medical Sciences, Cornell University

Master’s Project: Capstone

Dr. Xi Kathy Zhou

October 23, 2024

**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, race, BMI category, 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**

All datasets used for this analysis are available for download from the cBioPortal for Cancer Genomics database. For our research, we wanted to focus on studying any potential associations between race, sex, BMI, or smoking and tumor mutation burden or fraction genome altered in patients with varying forms of cancer. Tumor mutation burden is always measured in number of mutations per megabase but methods vary but what sequencing focuses on. Fraction genome altered is the percentage of a tumor’s chromosome regions that have been altered. For the dataset we created that would include all variables of interest (Race, Sex, BMI, and Smoking) we chose to merge 10 studies containing around or over 100 participants that had a BMI variable, the least studied, as well the other variables of interest for the most part. This final dataset included the “Bladder Urothelial Carcinoma (TCGA, Firehose Legacy)”, “Glioblastoma (CPTAC, Cell 2021)”, “Colorectal Cancer (MSK, J Natl Cancer Inst 2023)”, “Esophagogastric Cancer MSK 2024”, “Hepatocellular Carcinoma (MSK, 2024)”, “Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA/Firehose Legacy)”, “Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021)”, “Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy)”, “Skin Cutaneous Melanoma (TCGA, PanCancer Atlas)”, and “Uterine Corpus Endometrial Carcinoma (TCGA, Firehose Legacy)” datasets. Our process beganc by downloading the full folder for each dataset and reading the clinical patient as well as clinical sample files to RStudio.

The Bladder Urothelial Carcinoma (TCGA, Firehouse Legacy) dataset is from the GDAC (Georgia Data Analytics Center) Firehose which was previously known as TCGA Provisional. It contains data from 412 patients with Bladder Cancer as part of the Broad Institute of MIT and Harvard Firehose initiative. They primarily use whole exome sequencing to measure tumor mutation burden. This study used the smoking indicator values from the National Cancer Institute (NCI) Common Data Element (CDE) to standardize. The “1” represented “Lifelong Non-Smoker”, “2” represented “Current Smoker”, “3” represented “Current Reformed Smoker for > 15 yrs”, “4” represented “Current Reformed Smoker for < or = 15 yrs”, and 5 represented “Current Reformed Smoker, Duration Not Specified”. We decided the best course of action was to remain consistent with the other datasets that did not use this system and instead had only 3 categories, which were “Never”, “Current”, and “Former”. Any other changes to this dataset to aid in the merging process were to remove redudancies. For example “ASIAN” and “Asian” were combined in the race category, “Lymph node only” was changed to “Lymph node” for metastatic, and any cells with “[Not Available]” were treated as “NA”. We also created a BMI category in this and every dataset after based on the CDC official adult BMI categories – under 18.5, 18.5 to 25, 25 to 30, and 30+.

The Glioblastoma (CPTAC, Cell 2021) dataset was generated by the CPTAC (Clinical Proteomic Tumor Analysis Consortium) and used in the “Proteogenomic and metabolomic characterization of human glioblastoma” study. It contains data on 99 patients with the most aggressive and common brain tumor. Tumor mutation burden was calculated with number of mutations per megabase of sequenced DNA, focused on the protein coding regions. Again any redundancies were dealt with in data cleaning such as “Hispanic or Latino” and “HISPANIC OR LATINO” for the ethnicity variable or any odd cases where NA was not recognized in R. Given the size of this dataset, we opted to label all observations for tumor site as “Brain” over having multiple smaller categories. This dataset also used the NCI CDE for smoking indicator/history, which was changed in similar fasion to “Never”, “Current”, and “Former”. Finally, the primary tumor site variable was mutated to “Brain” for all locations of Glioblastoma in the brain. Otherwise, there would be too many groups in analysis with not enough samples/patients per group.

Colorectal Cancer (MSK, JNCI 2021) collected 1,516 samples from 818 patients with early onset and 698 patients with average onset colorectal cancer. This study “A Comprehensive Comparison of Early-Onset and Average-Onset Colorectal Cancers” looked into clinical and genomic differences between the two forms. TMB was measured again by total number of mutations per megabase of exome or targeted regions sequenced. During the first part of the data cleaning process, we dealt with the not available and redundancies first, such as any Asian or Native American related Race category getting combined into “Asian” or “American Indian or Alaska Native”. The same was done for “Adrenal Gland”, “Liver”, “Lymph Node”, “Peritoneum”, “Brain”, “Abdomen” (including abdominal wall), “Pelvis”, and “Colon” (all parts of the colon) for the metastatic site variable.

Esophagogastric Cancer (MSK, J Natl Cancer Inst 2023) dataset was used in the “Clinical and molecular characteristics of early-onset vs average-onset esophagogastric cancer” to determine significant differences seen between the 219 patients with early onset and 904 patients with average onset. 902 patients were included in the downloadable dataset from cBioPortal. TMB is the number of mutations per megabase of the sequenced exome. As above, we repeat the not available and redundancy corrections. This includes “WHITE” and “BLACK OR AFRICAN AMERICAN” for Race as well as “NOT HISPANIC OR LATINO” and “HISPANIC OR LATINO” for Ethnicity. For metastatic site, the redundant groups became “Adrenal Gland”, “Lymph Node”, “Peritoneum”, “Abdomen”, “Colon”, and “Spine”. To combate any confusion, all patients were identified as having Esophagogastric cancer in the “CANCER\_TYPE” column since that is what was being studied.

Hepatocellular Carcinoma (MSK, 2024) contains data on 1,370 patients with Hepatobiliary Cancer collected via the Memorial Sloan Kettering Cancer Center Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). This MSK-IMPACT tumor sequencing test is offered to MSK patients to detect mutations and adjust treatment care. The dataset was the only one that had did have a folder download and needed to be directly downloaded from the cBioPortal website. Thus, it was the first to include Fraction Genome Altered (not available in the patient/sample files in the folders). The information is also used in research. For TMB, the targeted sequencing method is used on genes known to be relevant in cancer. For data cleaning, we continue the process. Race combines the redundant groups to be “BLACK OR AFRICAN AMERICAN” and the ethnicity category “HISPANIC OR LATINO” includes all individual country subcategories or other hispanic terms. Primary tumor site combined groups include “Bile Duct” and “Gallbladder” while metastatic site included “Liver”, “Lymph Node”, “Peritoneum”, “Spine”, and “Abdomen”.

The Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA/Firehose Legacy) dataset is from the GDAC Firehose with information on 310 carcinomas from 308 patients. The cleaning system remains and the only redundant groups were for Race which was changed to “Asian”. This study used NCI CDE for smoking indicator so these values were updated to “Never” (“1), “Current” (“2”), and “Former” (any other values). The Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021) data targeted 412 intrahepatic cholangiocarcinoma tumors to study outcomes. TMB was measured using whole exome sequencing. According to the results of this original study, the “Genetic Determinants of Outcome in Intrahepatic Cholangiocarcinoma”, genetic profiling could help identify patient subgroups with poor outcome. The cleaning for this set was needed for the not available cells and conforming the smoking status to “Current”, “Former”, and “Never”. Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy) is another dataset from the GDAC Firehose which contains clinical data from 293 carcinomas in 292 patients with kidney cancers. For cleaning this dataset, we ensured the missing dating was properly recorded again and that the Race category for “Asian” was not repeated. In addition, since the NCI CDE was used for smoking, we updated this labels.

The Intrahepatic Cholangiocarcinoma (MSK, Hepatology 2021) dataset includes information on targeted sequencing of 412 intrahepatic cholangiocarcinoma tumor and normal sample pairs. It is used in the “Genetic Determinants of Outcome in Intrahepatic Cholangiocarcinoma” study that sequenced primary tumors to define associations among genetic alterations, clinicopathological variables, and outcome. Gene alterations were factors and clinical or pathologic variables, disease stage, and treatment were controlled for. We will use the dataset to investigate differences in tumor mutation burden and fraction genome altered between different sexes, races, and BMI values or categories. The whole exome sequencing method was used to calculate tumor mutation burden here as well. The data cleaning at this stage comprised of the steps previously taken – ensuring that there are no repeats of a category label that contains the same meaning and creating the BMI category variable.

The next dataset to be used was the Kidney Renal Papillary Cell Carcinoma (TCGA, Firehose Legacy) from GDAC Firehose as well. It studies this specific type of kidney cancer and includes whole exome sequencing, demographics with diverse patients, and information on tumors, treatments, and outcomes. Researchers use this data to identify genetic mutations or other features that could serve as a biomarker for diagnosis or treatment as well as comparison with other forms of kidney cancer. In our investigation, sex, race, and BMI values or categories post variable creation can be studied from this dataset. Data cleaning process remains the same for all data sets at this stage.

The Skin Cutaneous Melanoma (TCGA, PanCancer Atlas) data has been used in many publications such as “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”, “Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Piplines”, or “Genomic and Functional Approaches to Understanding Cancer Aneuploidy”. It was collected via the Pan-Cancer Atlas, which was from The Cancer Genome Atlas (TCGA), which aims to answer the big questions about cancer through analysis of TCGA datasets that profile human tumors. Data cleaning included the previous changes as well as combining redundant group names for “Bone”, “Ileum”, “Lung”, “Lymph Node”, “Vagina”, “Small Intestine”, “Small Bowel”, “Skin”, and “Back” for Metastatic sites. For the primary sites, the repeated groups were “Extremities”, and “Trunk|Unknown”, “Extremities|Trunk”. The final dataset we used in this combination process was “Uterine Corpus Endometrial Carcinoma TCGA/Firehose Legacy” data on 548 patients and also from the GDAC Firehose. The only cleaning needed here was for the “Asian” repeat and any not available groups that shoud be missing data again.

The final dataset used to analyze BMI was the Uterine Corpus Endometrial Carcinoma (TCGA, Firehose Legacy) which is also from the GDAC Firehose. It has been used to study the common gynecological cancer using genomic, transcriptomic, methylation, and clinical data from a diverse group of patients all diagnosed with endometrial carcinoma. This dataset, via whole genome sequencing TMB, can be applied to various forms of research – such as biomarker discovery, subtype investigation, or targeted treatments. For our research purposes, we can utilize this to study differences between BMI categories, sexes, and races in our variables of interest.

After the initial cleaning, the patient and sample data were combined into one dataset for each study, merging by patient ID. Each patient ID contained one row sample of unique data. Then, these dataframes were merged the same way with fraction genome altered from directly downloaded files for the studies from cBioPortal. These were finally all put together via rbind as all column names were consistent. The end product dataset contained patient ID (PATIENT\_ID), sex (SEX), race (RACE), ethnicity (ETHNICITY), age (AGE), overall survival status (OS\_STATUS), overall survival months (OS\_MONTHS), cancer type (CANCER\_TYPE), primary tumor site (SITE\_OF\_TUMOR\_TISSUE), metastatic sites (METASTATIC\_SITE\_PATIENT), smoking history (SMOKING\_HISTORY), smoking pack years (SMOKING\_PACK\_YEARS), BMI (BMI), BMI categories (BMI\_CATEGORIES), study (Study), tumor purity (TUMOR\_PURITY), tumor mutation burden (TMB), fraction genome altered (FGA), and the number of samples for each patient (Code\_frequency). The Code\_frequency column was created to ensure all samples were taken into account and checked for varying results in any of the variables.

The final steps of the dataset creation are to combine all categories that make the most sense to combine conceptually and statistically to ensure proper analysis and results. The race, sex, and smoking factors were previously dealt with by removing reduncies. Cancer type also did not need further combinations. However, primary tumor site and metastatic site needed more merging of categories to have significant results. The number of primary tumor site subcategories were not as high as the metastatic sites so they remained, for the most part, in their more specific categories. Some were combined into Colon (from left and right colon), Bile Duct, Skin, and Unknown (if unknown or not listed/specified) with all others remaining as is from the initial redundancy cleaning. The metastatic site was more involved. The large number of subcategories were combined into Abdomen, Arm, Bone, Chest, Female Reproductive, lymph, spine, thigh, trunk (referring to skin), and missing (which was the majority). However, due to the high amount of unknown data and unevenness of groups, this variable was disregarded in most later analysis.

Given that gender may be a potential factor influencing Tumor Mutation Burden (TMB) and Fraction Genome Altered (FGA), we aimed to ensure that the dataset used for subsequent analyses and statistical modeling contains sufficient gender information. To achieve this, we selected studies that recorded patient gender and consolidated the databases from these studies into a unified dataset that guarantees the inclusion of gender data (hereafter referred to as the ‘gender dataset’). The dataset contains 10,336 patients and 10,946 clinical observation records, including the following variables: patient\_id, sample\_id, cancer\_type, FGA, mutation\_count, number\_of\_samples\_per\_patient, age, cancer\_type, sex, TMB, primary\_tumor\_site, cancer\_type, and race\_category. The data were derived from the following studies: MSK Pancancer Study (MSK, Cell 2017), MSK MetTropism (MSK, Cell 2021), glioma\_mskcc\_2019 (MSK, Clinical Cancer Research 2019), Cancer Cell Line Encyclopedia (Broad, 2019), Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012), Bladder Cancer (MSK, Cell Reports 2022), Colorectal Cancer (MSK, JNCI 2021), Breast Cancer (MSK, Cancer Cell 2018), Breast Invasive Carcinoma (TCGA, PanCancer Atlas), Esophagogastric Cancer (MSK, Journal of the National Cancer Institute 2023), Metastatic Non-Small Cell Lung Cancer (MSK, Nature Medicine 2022), Metastatic Colorectal Cancer (MSK, Cancer Cell 2018) , Pan-Lung Cancer (TCGA, Nature Genetics 2016), Breast Invasive Carcinoma (TCGA, Firehose Legacy), Colorectal Cancer (MSK, JNCI 2021), Prostate Adenocarcinoma (MSK, European Urology 2020), and Prostate Adenocarcinoma (MSK, Clin Cancer Res. 2022). This interim dataset played a significant role as a transitional phase in the development of our final dataset.

Next, we created a dataset using all of the known studies from the previous BMI dataset, the other that centered around gender, and a new 24,146 sample “Cancer Therapy and Clonal Hematopoiesis” dataset. The new dataset contains clonal hematopoeisis mutations identified in blood samples from 24,146 patients, analyzed with MSK-IMPACT. The data was used in “Cancer therapy shapes the fitness landscape of clonal hematopoiesis” to study the process that drives certain clones to cancer for the purpose of future cancer therapies. It contains genomic data including TMB calculated from targeted sequencing of genes associated with cancer, clinical data, and cohort demographics. Researchers can use this information to understand clonal evolution and therapeutic insights, as in the paper mentioned, or predictive modeling. The demographics include sex, race, and smoking status, which will be used in our analysis. This dataset was also filtered for only non missing sex and later race as well for comparison. For this smoking version, after updating the variable names to remain consistent, filtering the data to only include observations with non missing smoking status values, and removing redundancies, we are left with data from 9 studies.

The studies left include the Bladder Urothelial Carcinoma, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma, Colorectal Cancer, Glioblastoma, Intrahepatic Cholangiocarcinoma, Kidney Renal Papillary Cell Carcinoma, Cancer Therapy and Clonal Hematopoiesis, MSK-IMPACT Clinical Sequencing Cohort, and Pan-Lung Cancer. Since metastatic site had many unknowns and categories that likely led to its insignificance, we decide to focus on broader, yet still specific, primary tumor site and cancer type. We included 33 cancer types in our smoking data set, which are Breast Cancer, Mesothelioma Cancer,Endometrial Cancer, Non-Small Cell Lung Cancer , Hepatobiliary Cancer, Esophagogastric Cancer, Bladder Cancer , Ovarian Cancer, Uterine Sarcoma, Germ Cell Tumor, Colorectal Cancer, Cervical Cancer, Prostate Cancer , Soft Tissue Sarcoma, Thyroid Cancer, Head and Neck Cancer, Anal Cancer, CNS Cancer, Bone Cancer, Kidney, Cancer of Unknown Primary, Pancreatic Cancer, Skin Cancer, Sex Cord Stromal Tumor, Peripheral Nervous System Cancer, Blood Cancer, Appendiceal Cancer, Adrenal Gland, Penile Cancer, Ampullary Cancer, Small Cell Lung Cancer, Gestational Trophoblastic Disease, Vaginal Cancer, Retinoblastoma, Histiocytosis, and Liver Cancer.

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. We began with graphical analyses such as boxplots to visualize our data prior to running individual factor analysis testing. In our individual testing, we utilized t-tests, Analysis of Variance, Kruskal-Wallis, and Man-Whitney U tests. ANOVA (mean) and Kruskal-Wallis (median) tests were used in all cases with more than groups while the t-test and Mann-Whitney U test were used in cases of just two groups. These were the tests most appropriate for comparing means and medians across their respective number of groups. Median and mean was chosen for testing to give a more comprehensive story, especially if the data did not meet normality assumptions.

In the multivariable analysis, 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. Although the database used for modeling had undergone preliminary processing and merging, additional processing specific to the mixed-effects model was still required. Since our final dataset is derived from eight different selected studies, the variables measured in each study are not identical. In fact, the variables retained in our database represent the union of variables across the eight studies. 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. 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 30 iterations. This approach ensured that the imputed data retained as much of the original characteristics as possible while maintaining a sufficiently large dataset size.

A graph of different colored lines

Description automatically generated

After 50 iterations, the means and standard deviations of the variables stabilized, indicating that the imputation results had converged.

Next, we performed variable selection for inclusion in the model. For all continuous variables (Age, , and survival time), we calculated the correlation with TMB and FGA. For all categorical variables, we conducted separate Analysis of Variance (ANOVA) tests with TMB and FGA. In addition, we performed a multicollinearity check for all variables to ensure the independence of all predictor variables from one another. Ultimately, ‘METASTATIC’ was excluded due to strong collinearity with ‘Cancer Type.’ The final predictor variables retained in the model were GE, SEX, RACE, TUMOR PURITY, SMOKING\_HISTORY, PATIENT\_ID, Study, and CANCER\_TYPE.

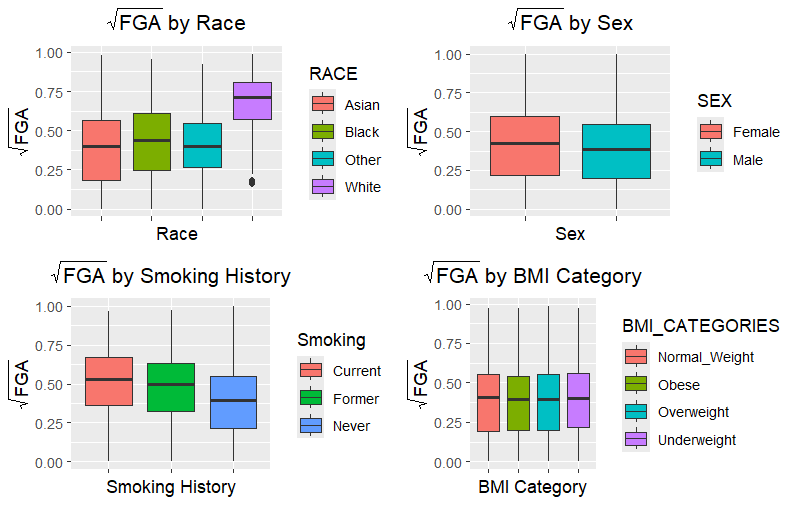
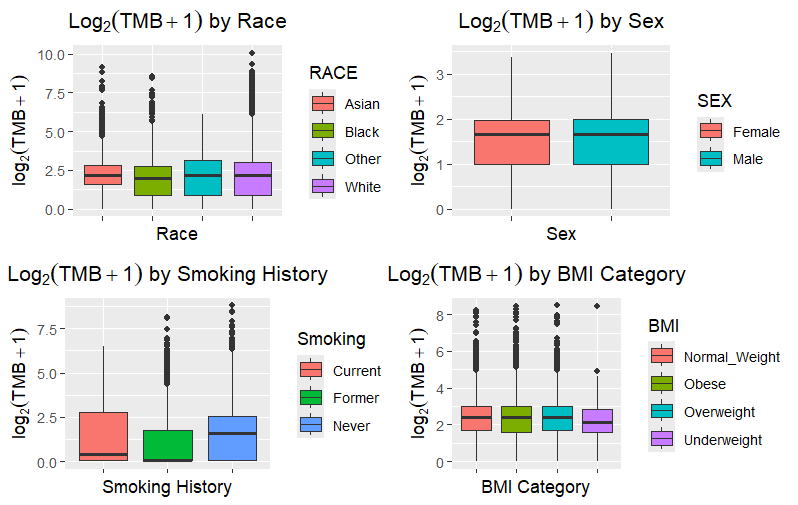
Since the distributions of the response variables, Fraction Genome Altered (FGA) and Tumor Mutation Burden (TMB), were not normally distributed, we transformed FGA to √FGA and TMB to log2(TMB+1) to meet the normality assumption of the linear mixed-effects regression model.

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.

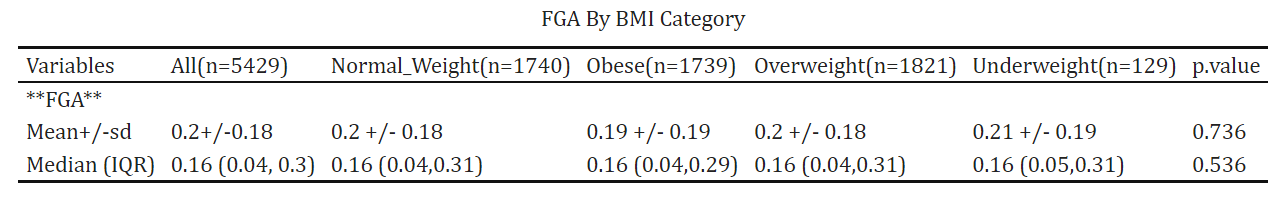
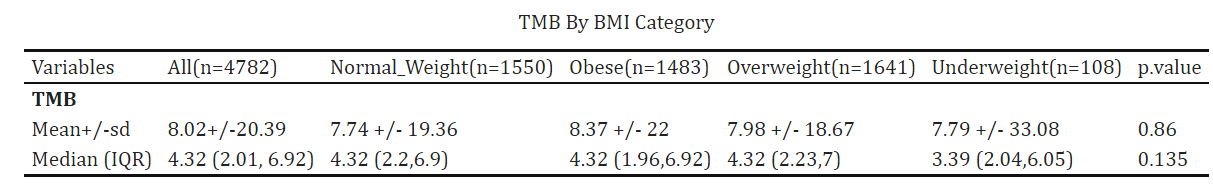
After completing the modeling process, we performed a normality test on the residuals. Since the database size exceeded the sample size limits of the Shapiro-Wilk test in the stats package, we assessed normality using a Q-Q plot of the residuals. Finally, we diagnosed and compared models’ goodness of fit using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the R-squared calculation function from the performance package.

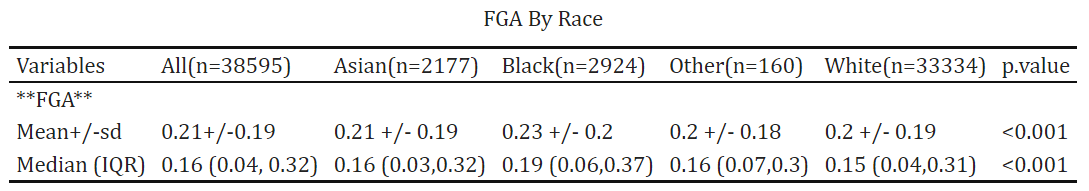
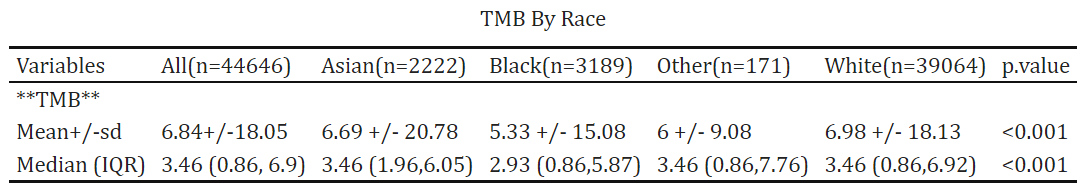
**Results**

The first part of our analysis was to have a better understanding of the data through descriptive statistics and individual factor analysis. We looked at how the data was spread throughout various groups in each of the dataset subsets, 4 datasets from all studies discussed in methods combined with the factor that is filtered for varying. We transformed the variables as appropriate in correspondence with our findings and the models. The data is not distributed in a form that can be studied well when it is not transformed. Most datapoints are concentrated around a certain value for tumor mutation burden, with not many other data points beyond outliers, and there was a very small range for fraction genome altered. In the dataset where there was no missing Race data, we do see much overlap in the values for log base 2 of the sum of TMB and 1 but clear differences in the square root of fraction genome altered. The “White” category did not overlap much with the other categories. For the dataset with no missing Sex values, there is a lot of overlap for both transformed TMB and FGA. While the dataset with no missing smoking history factor values does have overlap, there are also clear visual differences in the boxplots between the never and former or current categories for both variables. Last, the BMI categories appear to be the closest in spread and center from the boxplots for both transformed FGA and TMB.

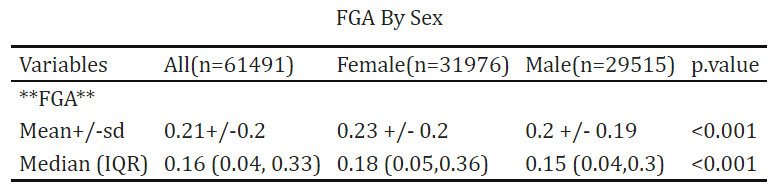
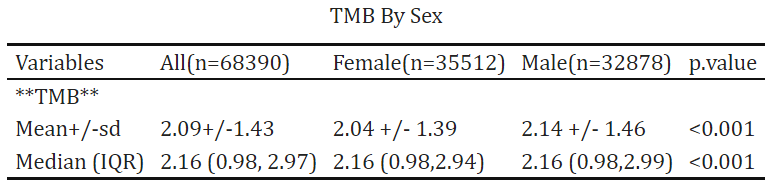


The BMI category was the first in our more in depth individual factor analysis. There were 4 groups with a total of 4,782 observations with no missing BMI and TMB data. This was made up of 1,550 (.324) within normal weight range, 1,483 (.31) in the obese category, 1641 (.343) overweight observations, and a much lower number of 108 (.023) for those classified as underweight. This is similar to the U.S population, though obese is slightly higher. The groups for the 5,429 no missing FGA and BMI were 1,740 (.32) for normal, 1,739 (.32) for obese, 1,821 (.335) for overweight, 129 (.024) for underweight. This was very similar to the TMB distribution. For both of these variables, the ANOVA test for means and Kruskal-Wallis for medians did not show significant results. As seen in the figure below, all p-values very above .1 so there was not evidence of diffference in mean or median between groups. Due to this, we decided BMI category was not necessary for our final model.

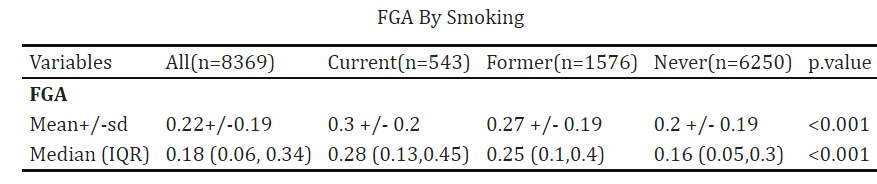
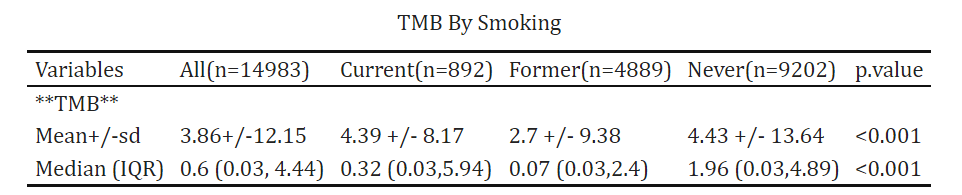
 The 44,646 non missing race and TMB dataset was made up of 2,222 (.05) Asians, 3,189 (.071) Black people, 39,064 (.875) White people, and 171 (.004) who identified as anything other than these categories. This is a lower percentage for Black people and “Other” as well as a higher population of White people than in the US. This is also not reflective of the world population, especially with such a low percentage within the Asian group. This is seen again in the 38,595 FGA and Sex dataset. About 5.6% of the observations were Asian people, 7.6% were Black people, 86.4% White people, and .4% identified as other. For both mean and median, there was evidence of statistically significant differences in both TMB and FGA. All p-values were less than .001 for these ANOVA/Kruskal-Wallis tests, as seen in the figure.



The non missing TMB and Sex dataset had 68,390 observations comprised of about 51.9% (35512) females and 48.1% (32878) males. These groups were fairly even. This is seen again with the 52% (31976) females and 48% (29515) males in the 61,491 FGA and Sex dataset. Again we see statistically significant p-values in the T-test and Mann-Whitney U test. This would suggest evidence of statistically significant difference in the mean and median between the groups.

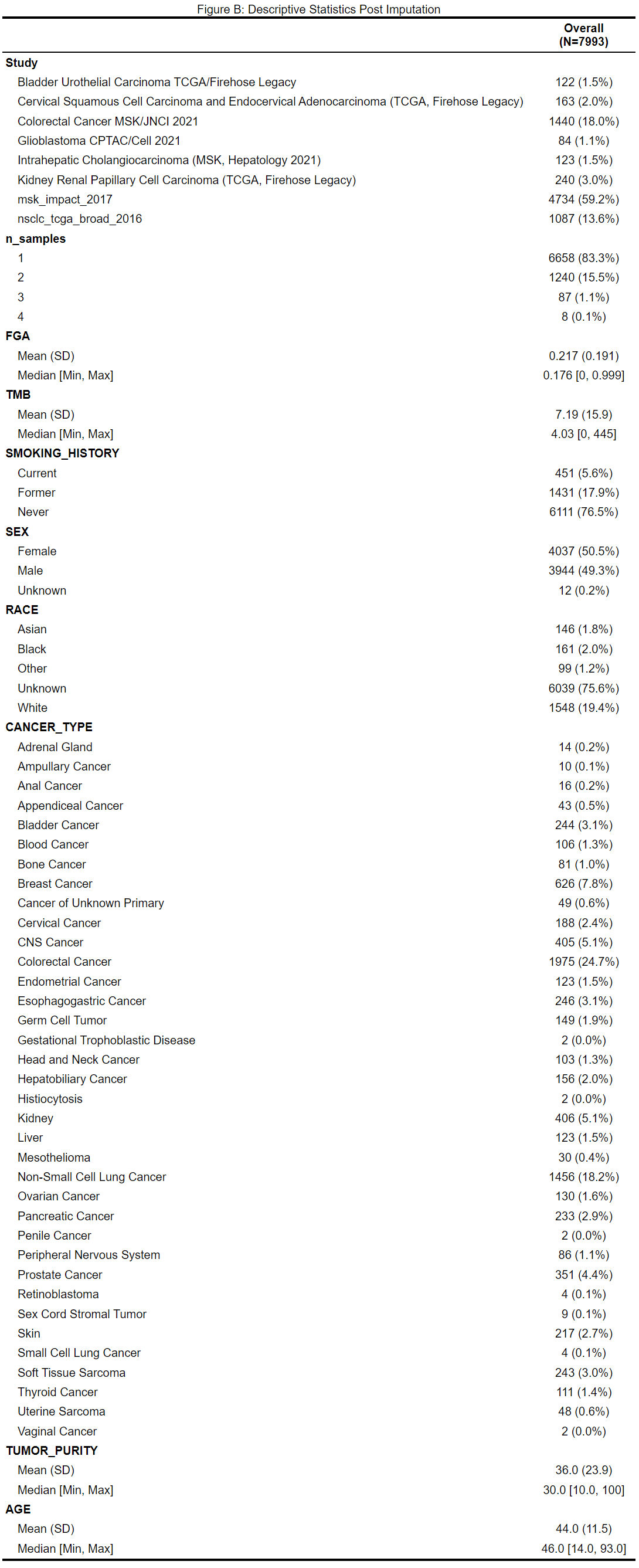
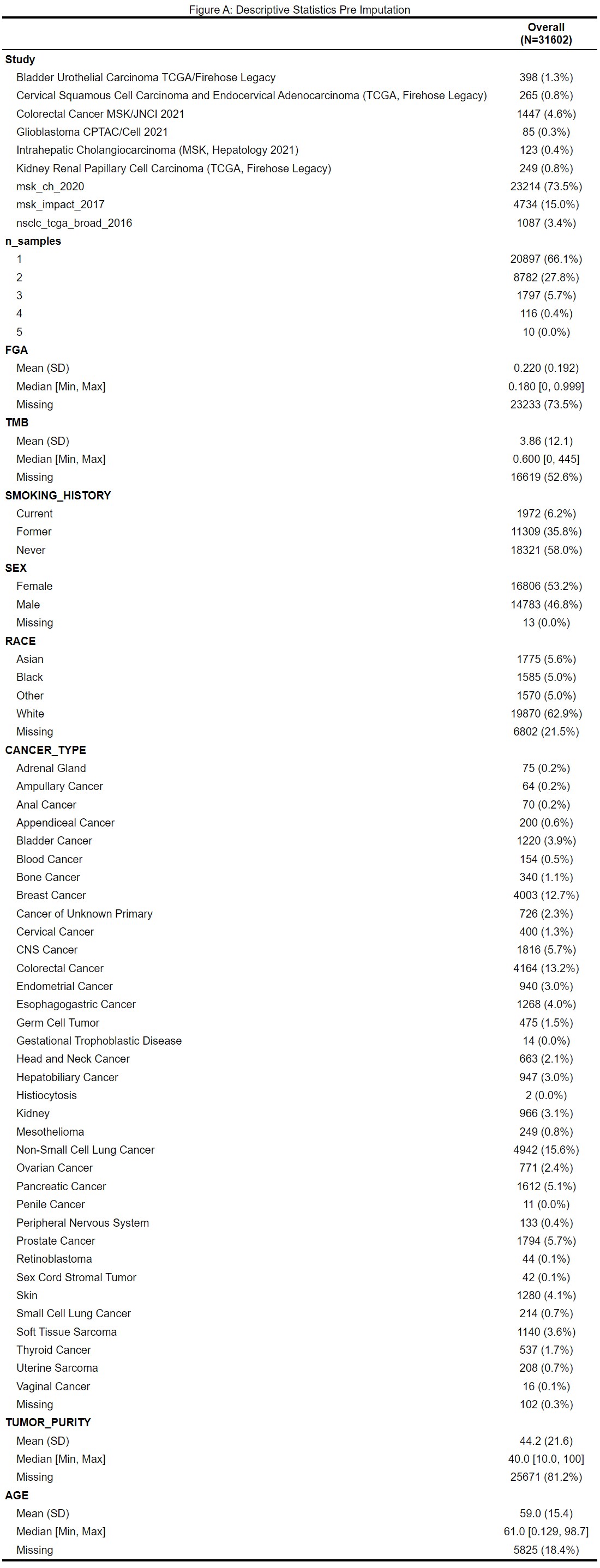


Smoking history contained 3 levels of current, former, and never. In the smoking history and TMB non missing dataset, there were 14,983 observations with 892 current smokers (.06), 1576 former smokers (.326), and 9,202 (.614) never. This is less than the US and worldwide percentage, according to the World Health Organization, for current smokers. We see this again for the 8369 observations in the FGA and smoking history non missing dataset. 6.5% are current smokers, 18.8% are former smokers, and 74.7)% are non-smokers. There appears to be statistically significant evidence of differences in both mean and median for both variables.



We decide to focus on the non missing smoking dataset as it contains all significant factors with 31,602 data points. Before beginning the model, we look at the general descriptive statistics for this non missing smoking dataset, shown in figure a on the left. Most of the data in this dataset is from Cancer Therapy and Clonal Hematopoiesis study with 23,214 samples, all including smoking history. For the most part, each patient only has 1 sample (66.1%) but there are a good portion with 2 (27.8%) and even ten patients had 5. Within our factors and variables of interest, we see similar distributions. The FGA does not have much range and TMB is very concentrated around 3.86 despite that large range. TMB and FGA have high (over 50%) missingness. The spread of smoking history has been discussed for this and the proportion of men and women remain very close. Though there are some differences as Asian/Black/Other people have a similar percentage of around 5% and the White people have a proportion now lower, the story of discrepancy with the general population continues. For the cancer types, we kept a list of 33 specific types for analysis, with the largest categories being “Non-Small Cell Lung Cancer”, “Colorectal Cancer”, and “Breast Cancer”. We will go on to further explore these variables with the models. Tumor purity had a very large amount of missing data (over 80%). Age would appear to be skewed left, with a center at about 61 that is impacted by some lower values. The impact of missingness is what led to imputation before modeling.

The figure B table shows the new distribution of variables post the necessary imputations. There is still a majority of one sample patients with no 5 sample patients anymore. The FGA values are of very similar distribution but the TMB values appear more normal looking post transformation now, with no missing data for either. The distribution of smoking history indicator shows a higher distinction between never smoking and former/current smoker, which does not follow real world demographics, especially for the older population. The age in this strongly centered around younger values of the early 40s. The race categories had an even stronger disparity between the White category and other categories, suggesting more missing values for those other categories. The tumor purity values remain right skewed with center at around 30. Finally, most patients are still from the Cancer Therapy and Clonal Hematopoiesis study with most common forms of cancer being colorectal and non-small cell lung cancer. This is the dataset to be used for the model construction.



For the model with the transformed TMB as the response variable, the formula is as follows:

*tumor*

*REML criterion at convergence: 22775.3*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-3.9720 -0.2679 -0.0178 0.2309 4.9955*

*Random effects:*

*Groups Name Variance Std.Dev.*

*PATIENT\_ID (Intercept) 0.8377 0.9153*

*CANCER\_TYPE (Intercept) 0.3275 0.5722*

*Study (Intercept) 0.2300 0.4796*

*Residual 0.2355 0.4852*

*Number of obs: 7993, groups: PATIENT\_ID, 7309; CANCER\_TYPE, 36; Study, 8*

*The model fitting results are as follows:*

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 1.559e+00 2.404e-01 2.100e+01 6.484 2.00e-06 \*\*\*

AGE 4.604e-03 1.105e-03 4.667e+03 4.166 3.16e-05 \*\*\*

SEXMale 8.068e-02 2.751e-02 7.304e+03 2.933 0.00337 \*\*

SEXUnknown 1.813e+00 3.016e-01 7.437e+03 6.013 1.91e-09 \*\*\*

RACEBlack 1.686e-02 1.043e-01 3.836e+03 0.162 0.87165

RACEOther 2.700e-02 1.124e-01 2.906e+03 0.240 0.81014

RACEUnknown 1.706e-01 1.608e-01 5.829e+02 1.061 0.28912

RACEWhite 5.506e-02 7.616e-02 3.253e+03 0.723 0.46974

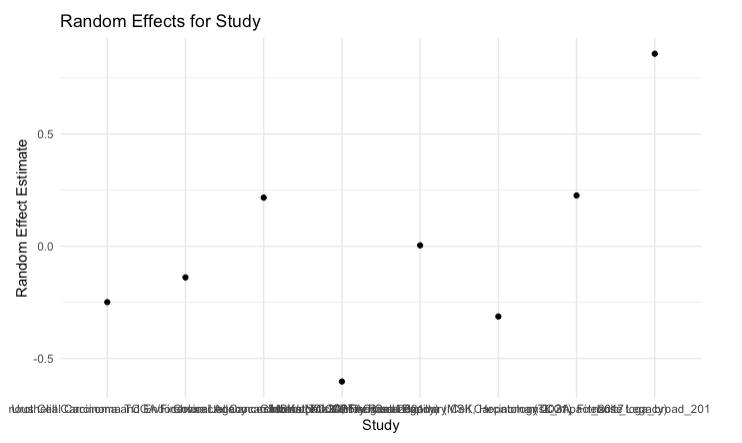
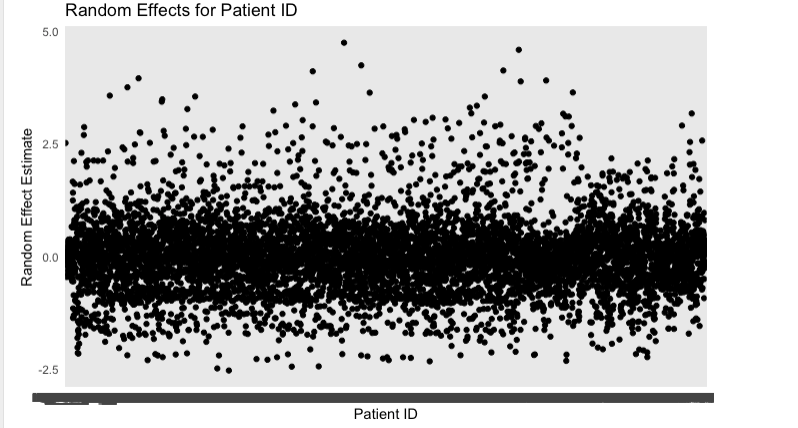
TUMOR\_PURITY 2.707e-03 5.915e-04 7.098e+03 4.577 4.80e-06 \*\*\*

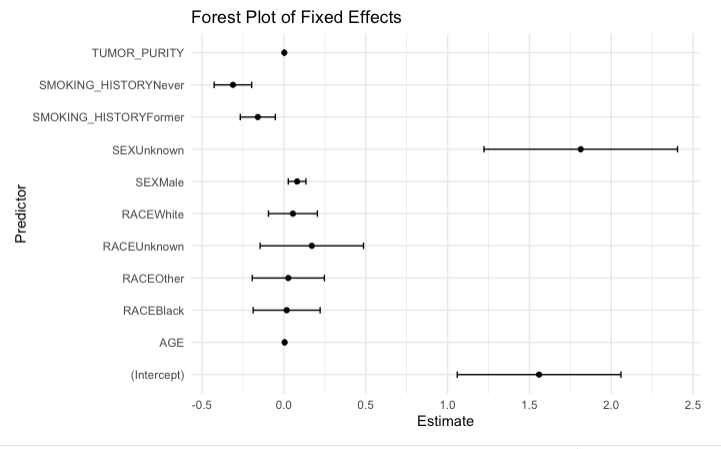
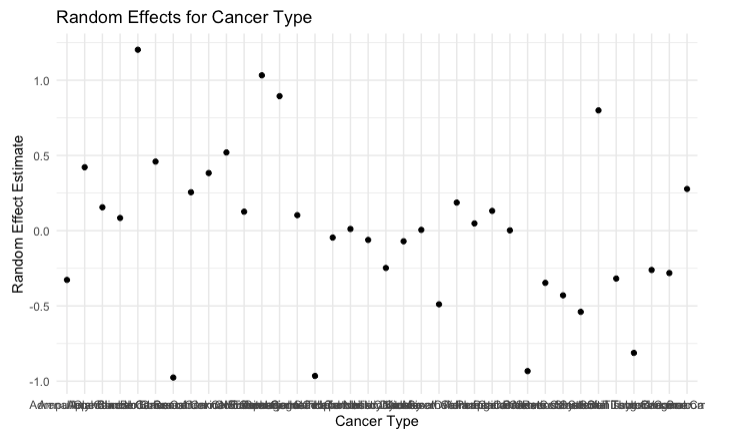
SMOKING\_HISTORYFormer -1.594e-01 5.459e-02 7.833e+03 -2.919 0.00352 \*\*

SMOKING\_HISTORYNever -3.113e-01 5.843e-02 7.361e+03 -5.327 1.03e-07 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

For the random effects section, the variance of **PATIENT\_ID (Intercept)**is 0.8377, with a standard deviation of 0.9153, indicating a significant random intercept effect on transformed TMB at the patient level.Th variance of **CANCER\_TYPE (Intercept)** is 0.3275, the variance of **Study (Intercept)** is 0.2300, indicating a random effect at the study level and at the cancer type level.





For the fixed effect section, the estimated coefficient of age is 4.604e-03, with a p-value < 0.001, indicating that age has a significant positive effect on TMB. Compared to the baseline (female), the estimate for males is 8.068e-02, with a p-value of 0.00337, suggesting that gender has a significant effect on TMB. The estimated coefficient of tumor purtiy is 7.097e-02, with a p-value < 0.001, indicating that tumor purity significantly influences TMB. Compared to current smokers, the coefficient for former smokers is -1.594e-01, with a p-value of 0.00352, and the coefficient for people who never smoked is -3.113e-01, with a p-value of 1.03e-07, indicating a significant negative impact of non-smoking quitting-smoking on TMB.

In the diagnosis of the goodness of the fit, this model achieved a conditional R-squraed value of 0.857, indicating a good fit. An more complicated alternative model with an interaction between studies and cancer type structure:

*+ (1|Patient\_ID) + 1(Study:Cancer\_TYPE) + ε*

*We got result of model fit：*

*Random effects:*

*Groups Name Variance Std.Dev.*

*PATIENT\_ID (Intercept) 0.8375 0.9151*

*Study:CANCER\_TYPE (Intercept) 0.3548 0.5957*

*Residual 0.2356 0.4854*

There was no significant improvement in the explained variance of the random effects. Additionally, the alternative model’s AIC and BIC were 22,805.57 and 22,903.38, respectively, with a conditional R-squared of 0.837. Compared to the original model, these metrics did not show a substantial improvement. After conducting an ANOVA test comparing the two models, the Pr(>Chisq) value was 0.3667, indicating no significant difference. Therefore, increasing the complexity of the random effects structure does not improve model performance. With a Q-Q plot we can find the residual of this model is relative normal distribution.

A graph showing a line

Description automatically generated

For the model with the transformed FGA as the response variable, the formula is as follows:

*The model fit result:*

*Random effects:*

*Groups Name Variance Std.Dev.*

*PATIENT\_ID (Intercept) 0.032618 0.18061*

*CANCER\_TYPE (Intercept) 0.004663 0.06828*

*Study (Intercept) 0.009315 0.09651*

*Residual 0.012053 0.10978*

*Number of obs: 7993, groups: PATIENT\_ID, 7309; CANCER\_TYPE, 36; Study, 8*

*Fixed effects:*

*Estimate Std. Error df t value Pr(>|t|)*

*(Intercept) 4.745e-01 4.594e-02 1.940e+01 10.327 2.52e-09 \*\*\**

*AGE -5.256e-04 2.322e-04 5.432e+03 -2.264 0.023630 \**

*SEXMale 1.105e-02 5.584e-03 7.148e+03 1.979 0.047844 \**

*SEXUnknown -1.935e-01 6.154e-02 7.479e+03 -3.144 0.001675 \*\**

*RACEBlack -1.869e-02 2.205e-02 4.586e+03 -0.848 0.396539*

*RACEOther -3.245e-02 2.394e-02 3.543e+03 -1.355 0.175356*

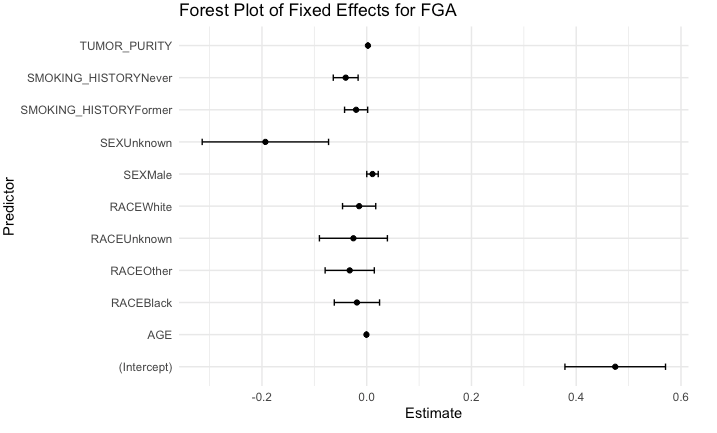
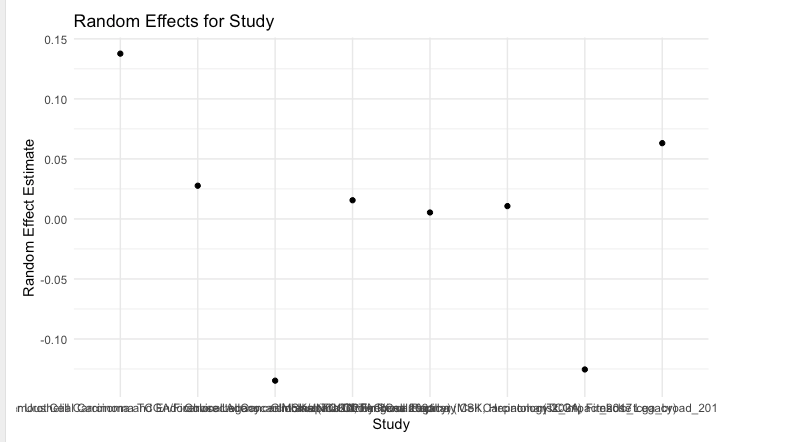
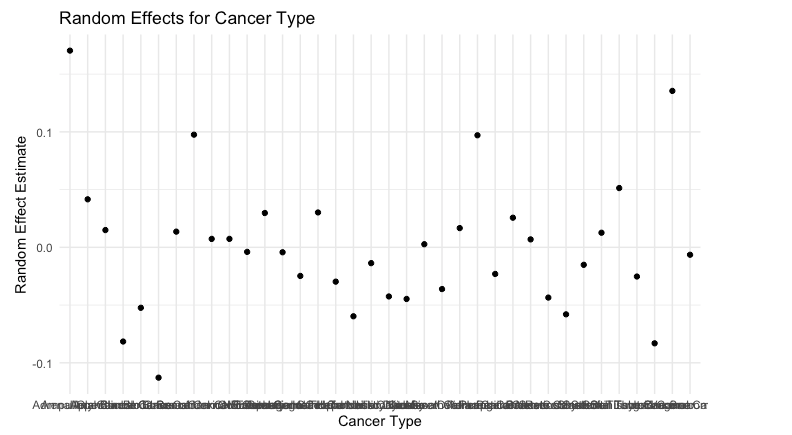
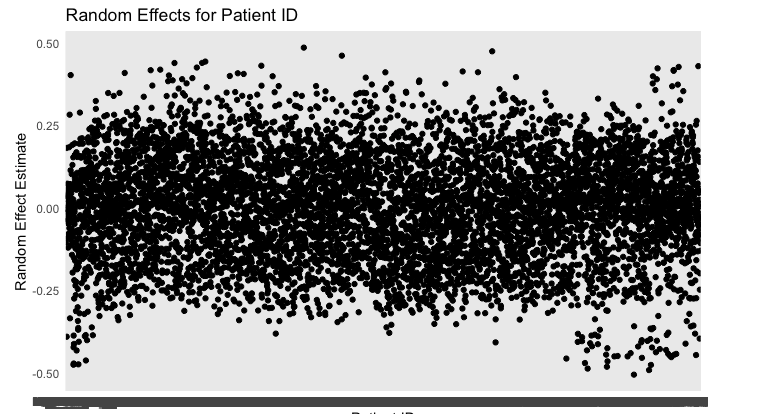
*RACEUnknown -2.540e-02 3.304e-02 5.551e+02 -0.769 0.442291*

*RACEWhite -1.445e-02 1.617e-02 3.948e+03 -0.894 0.371593*

*TUMOR\_PURITY 2.335e-03 1.223e-04 7.504e+03 19.086 < 2e-16 \*\*\**

*SMOKING\_HISTORYFormer -2.022e-02 1.125e-02 7.912e+03 -1.797 0.072365 .*

*SMOKING\_HISTORYNever -4.013e-02 1.210e-02 7.626e+03 -3.318 0.000912 \*\*\**



For random effect section: The patient ID, studies and cancer types all contributed a variation to the transformed FGA, but the studies’ impact is related small. For the fixed effect model, age has an estimated coefficient of -0.0005256 and p-value of 0.0236 indicate a slight but significant negative relationship between age and TMB; male has an estimated coefficient of 0.01105 and p-value of 0.0478, suggesting a small but significant positive effect of being male compared to females; TUMOR\_PURITY has an estimated coefficient of 0.002335 and a p-value less than 2e-16 indicate a highly significant positive relationship between tumor purity and FGA; Never smokers has an estimated coefficient of -0.04013 and p-value of 0.000912 compared to baseline (current smoker), showing a significant negative relationship with FGA, indicating that never smokers have lower FGA compared to others; former smokers has an estimated coefficient of -2.022e-02, but the impact is not siginificant due to a large p-value of 0.072365. We did not find that race has a significant impact of tranformed FGA. The model achieved a conditional R-squared value of 0.804, indicating a good fit. Based on the Q-Q plot diagnostic, the model residuals appear to follow a normal distribution.

A graph of a normal q-q plot

Description automatically generated

We also tested an alternative model that introduced a nested random effects structure:

*performance::r2(mixed\_model\_FGA\_n3)*

*# R2 for Mixed Models*

*Conditional R2: 0.789*

*Marginal R2: 0.050*

*Models:*

*mixed\_model\_FGA\_x: sqrt(FGA) ~ AGE + SEX + RACE + TUMOR\_PURITY + SMOKING\_HISTORY + (1 | PATIENT\_ID) + (1 | Study) + (1 | CANCER\_TYPE)*

*mixed\_model\_FGA\_n3: 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 15 -2565.4 -2460.6 1297.7 -2595.4*

*mixed\_model\_FGA\_n3 15 -2567.2 -2462.4 1298.6 -2597.2 1.8067 0*

Similar to the TMB model, after fitting and conducting an ANOVA test, no significant improvement in model performance was observed with the introduction of nested random effects. To investigate whether smoking history has varying degrees of influence on Tumor Mutation Burden across different cancer types, we constructed a new model:

*And got the result below:*

*Random effects:*

*Groups Name Variance Std.Dev. Corr*

*PATIENT\_ID (Intercept) 0.82190 0.9066*

*CANCER\_TYPE (Intercept) 0.35618 0.5968*

*SMOKING\_HISTORYFormer 0.01007 0.1004 -0.52*

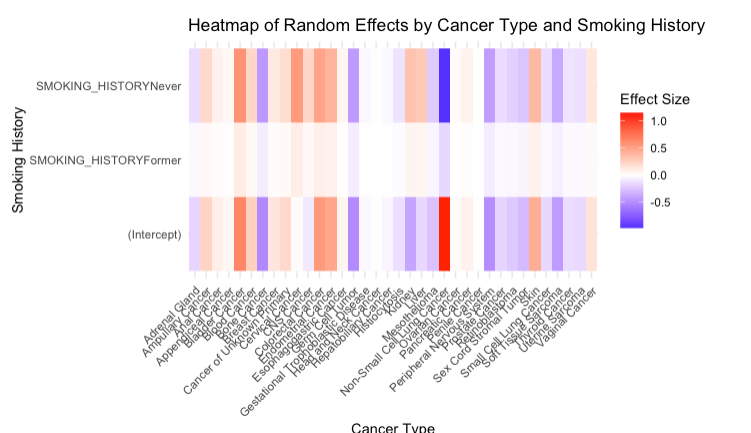
*SMOKING\_HISTORYNever 0.33009 0.5745 -0.53 1.00*

*Study (Intercept) 0.09676 0.3111*

*Residual 0.23251 0.4822*

We found that across different cancer types, the effect of former smokers on transformed TMB was relatively similar to that of current smokers. In contrast, the impact of never smokers on the response variable showed greater variation.

A graph of smoking and cancer

Description automatically generated

The figure illustrates the differences in the effect levels of never smokers and former smokers, compared to current smokers, on transformed TMB across different cancer types. It is obvious that the effect of smoking history varies greatly across different cancer types. Notably, never\_smoking has a particularly pronounced negative effect on transformed TMB in Small Cell Lung Cancer.

**Discussion**

The primary objective of this study is to investigate the impact of different cancer types (CANCER\_TYPE), patient characteristics (AGE, SEX, RACE), tumor purity (TUMOR\_PURITY), and smoking history (SMOKING\_HISTORY) on Tumor Mutation Burden (TMB) and Fraction Genome Altered (FGA) by constructing mixed-effects models. 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.

First, in the model incorporating random effects, we found that both individual patient differences and cancer type play significant roles in the variation of genomic alterations. The random effect of patient ID demonstrated substantial differences in TMB and FGA within the same cancer type, indicating considerable variability in genomic alterations across individuals. Additionally, the random effect of cancer type was highly significant, suggesting systematic differences in genomic alterations across various cancer types.

Further analysis of smoking history revealed that its effect on TMB varies by cancer type. Through the random slope model, we captured an opposite directional influence of ‘never smoked’ and ‘former smoker’ compared to the baseline (current smoker) in certain cancer types. This indicates that smoking history is not only a fixed influencing factor but also exhibits a high degree of heterogeneity in its effects across different cancer types.

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 0.837 (for TMB) and 0.804 (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.

We also found that increasing age, male gender, higher tumor purity, and current smoking status are significant factors that promote genomic alterations. Smoking history’s effect on TMB is significantly moderated by cancer type, with the direction and magnitude of the impact varying notably across different cancer types. Additionally, the significant individual differences, though may be due to the large sample size given the small effect, among patients highlight the importance of continuing to prioritize personalized treatment to optimize clinical outcomes. Future studies should focus on more diverse generalizable groups to have more effective analyses.

**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

‌