Assignment2 : final exam

2018250106 O Jeong min

#Topic : Correlation between APOBEC signatures and stage according to age in nonsmokers

#1. Background - APOBEC APOBEC is a kind of cytidine deaminases. The enzyme APOBEC which is found in cancer cells help to increase resistance to the drugs by rapidly changing cancer cells genetically. APOBEC also helps to protect against viral infections of mammals. When mismanifested, APOBEC enzyme can be a major factor of mutations in many kinds of cancer.(Shi, Ke; et al., Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B, 2017.)

#2. Load dataset Load necessary “library”s and data from the paper.

```
library(readxl)
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────────────────────────────────── ti
dyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────────────────────── tidyver
se_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(pander)
```

```
#readxl::excel_sheets('Chen2020/1-s2.0-S0092867420307431-mmc1.xlsx')
#readxl::excel_sheets('Chen2020/1-s2.0-S0092867420307431-mmc6.xlsx')

d1_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-mmc1.xlsx', sheet=2, na ="NA")
d6_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-mmc6.xlsx', sheet=2, na ="NA")
```

As a result of loading the data file, it was determined that “Correlation between APOBEC signatures and stage according to age in nonsmokers” could be recognized using d1_2 data and d6_2. After examining the correlation between the age and tumor stage of lung cancer patients in non-smokers using the data of d1_2, we compare the disease pattern between patients with APOBEC characteristic and those without APOBEC using the data of d6_2.

#3. Correlation between stage and age in nonsmokers

#3.1. Transforming data Make data into more useful form. Change the column name so that no spaces are included. Use “filter” to remain information that contains only non-smokers data.

```
colnames(d1_2)[5]<-"Smoking_Status"
pander(head(d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke")))
```

Table continues below

ID	Proteome_Batch	Gender	Age	Smoking_Status	Histology Type	Stage	EGFR_Status	Primary Tumor Location
P002	B01-2	Male	73.78	Nonsmoke	ADC	IB	others	LUL
P004	B01-4	Female	52.98	Nonsmoke	SCC	IA	exon19del	RLL
P006	B02-2	Female	46.86	Nonsmoke	ADC	IB	WT	RLL
P007	B02-3	Male	67.41	Nonsmoke	ADC	IIA	WT	RLL
P009	B03-1	Female	53.8	Nonsmoke	ADC	IIA	L858R	LLL
P010	B03-2	Female	56.48	Nonsmoke	ADC	IB	exon19del	LUL

Find the age distribution of the patient by finding the maximum and minimum values of the patient’s age.

```
d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% arrange(Age) %
>% select(Age) %>% max()
```

```
## [1] 85.86448
```

```
d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% arrange(Age) %
>% select(Age) %>% min()
```

```
## [1] 40.22724
```

Then using “mutate”function, create a new column “Ages” which represents informations organized in 10-year units. Designate this information as d1_2_Ages.

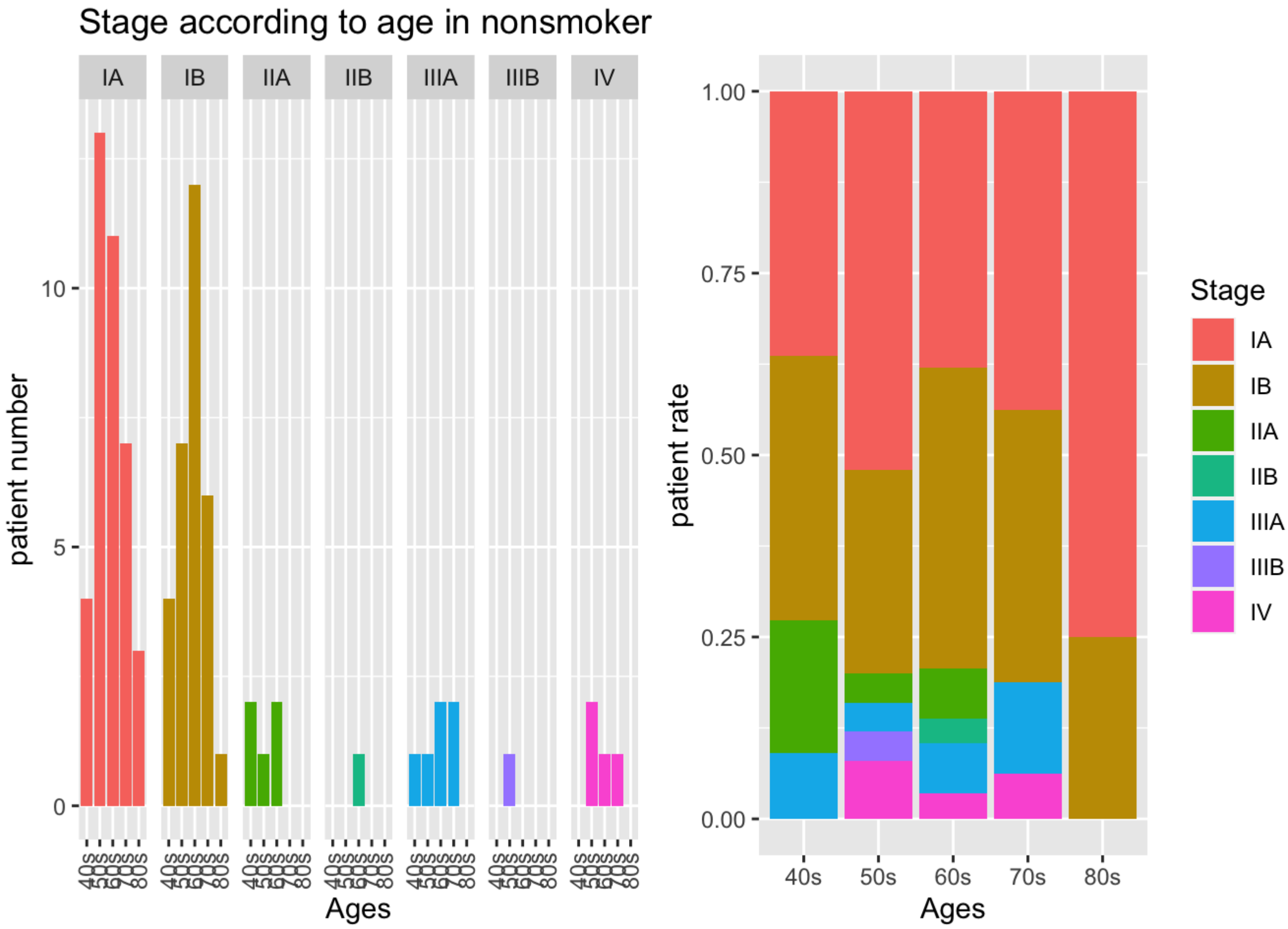
```
d1_2_Ages <- d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% mu
tate(Ages=case_when(Age<50 ~ "40s", Age<60 ~ "50s", Age<70 ~ "60s", Age<80 ~ "70s", A
ge>=80 ~ "80s"))
```

#3.2. Drawing plot Finally, draw the graphs that show the relationship between the tumor stage and the patient’s age in non-smokers.

```
pander(head(d1_2_Ages %>% group_by(Stage, Ages) %>% count()))
```

Stage	Ages	n
IA	40s	4
IA	50s	13
IA	60s	11
IA	70s	7
IA	80s	3
IB	40s	4

```
aa <- d1_2_Ages %>% select(Stage, Ages) %>% count(Stage, Ages) %>%
  ggplot(aes(Ages, n, fill=Stage)) +
  geom_bar(stat="identity", position="stack") +
  facet_grid(~Stage) +
  theme(legend.position = "none") +
  labs(title = 'Stage according to age in nonsmokers') +
  ylab("patient number") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

bb <- d1_2_Ages %>% select(Stage, Ages) %>% count(Stage, Ages) %>%
  ggplot(aes(Ages, n, fill=Stage))+
  geom_bar(stat="identity", position='fill') +
  ylab("patient rate") +
  labs(title = '')

grid.arrange(aa, bb, ncol = 2)
```



#3.3. Result and analysis In non-smoker lung cancer patients, stage IA and IB accounted for the most cases and the number of the patients was the smallest in the order of Ages 80s-40s-70s-50s-60s. It is presumed that the tumor stages in all patient Ages have a generally similar pattern. It can be inferred that stage IA and IB do not show specificity according to the age of the patient. The number of patient cases varies according to the Ages of the patient, and it may be thought that the incidence of cancer varies with age. But in my opinion, that thought is an incorrect analysis. We can guess that it is more correct to think this way: the reason why the number of 80s cases is small is because older patients are in risks of death by tumors.

#4. Correlation between APOBEC signatures and stage according to age in nonsmokers

#4.1. Transforming data This time, we use data “d6_2. Filter some useful data. Delete unnecessary columns. To utilize the data, change the order of rows and columns of the dataframe and rename the column. Then mutate data to make”Ages" column. Designate this dataframe as “d6_2_data”.

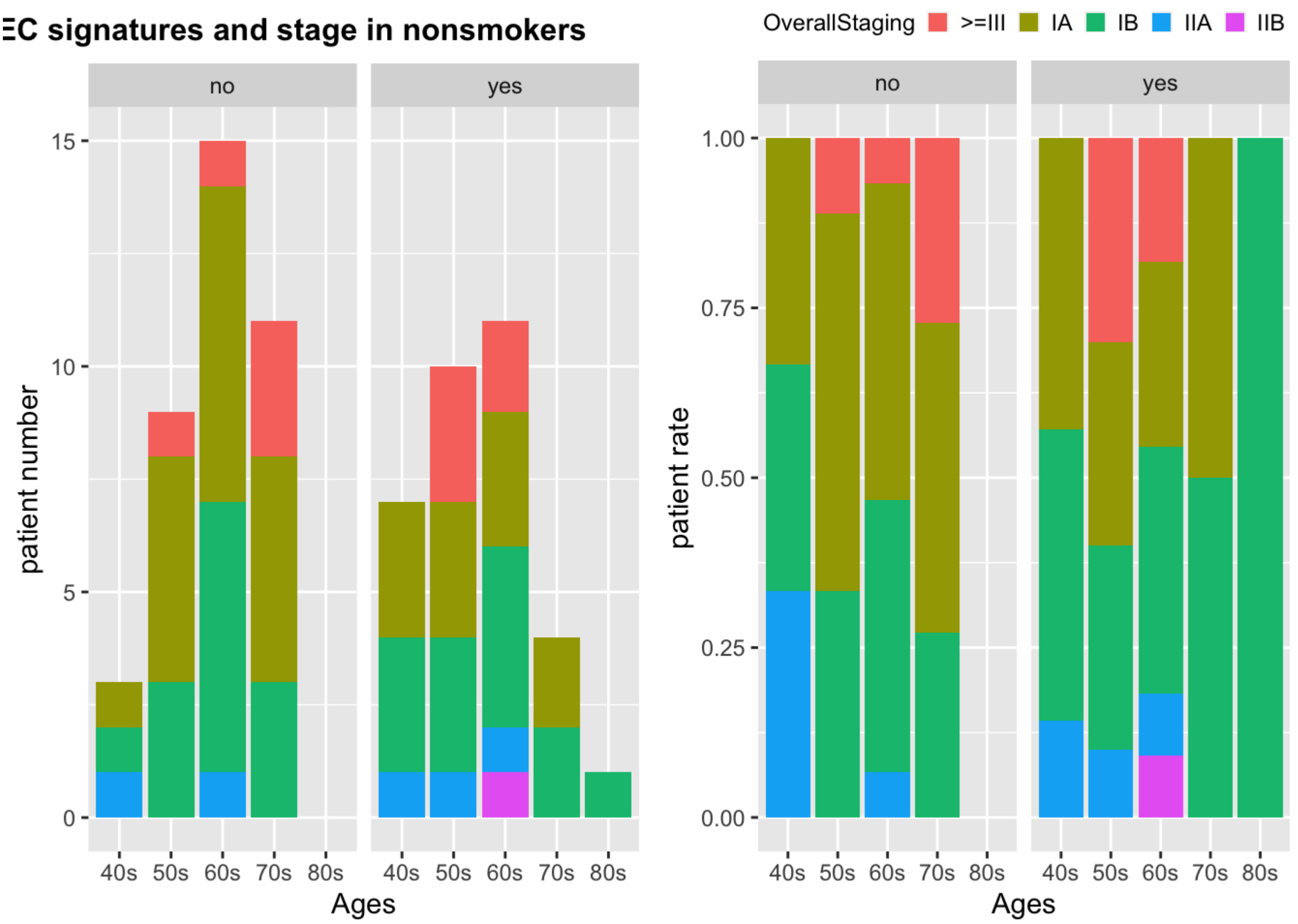
```
d6_2_data <- d6_2[,-c(2,3,4,5)] %>% filter(id %in% c("Age", "SmokingStatus", "Overall
Staging", "APOBEC_Enriched"))
d6_2_data <- d6_2_data %>% t() %>% as.data.frame()
d6_2_data <- rownames_to_column(d6_2_data, var = "id")
d6_2_data <- rename(d6_2_data,Age = V1, SmokingStatus=V2, OverallStaging=V3, APOBEC_E
nriched=V4)
d6_2_data <- d6_2_data[-1,]
d6_2_data <- d6_2_data %>% filter(SmokingStatus=="no") %>% mutate(Ages=case_when(Age<
50 ~ "40s", Age<60 ~ "50s", Age<70 ~ "60s", Age<80 ~ "70s", Age>=80 ~ "80s"))
```

#4.2. Drawing plot Visualization the graph that shows the relationship between APOBEC signatures and stage according to age in nonsmokers.

```
pander(head(d6_2_data %>% group_by(APOBEC_Enriched) %>% count(OverallStaging, Ages)))
```

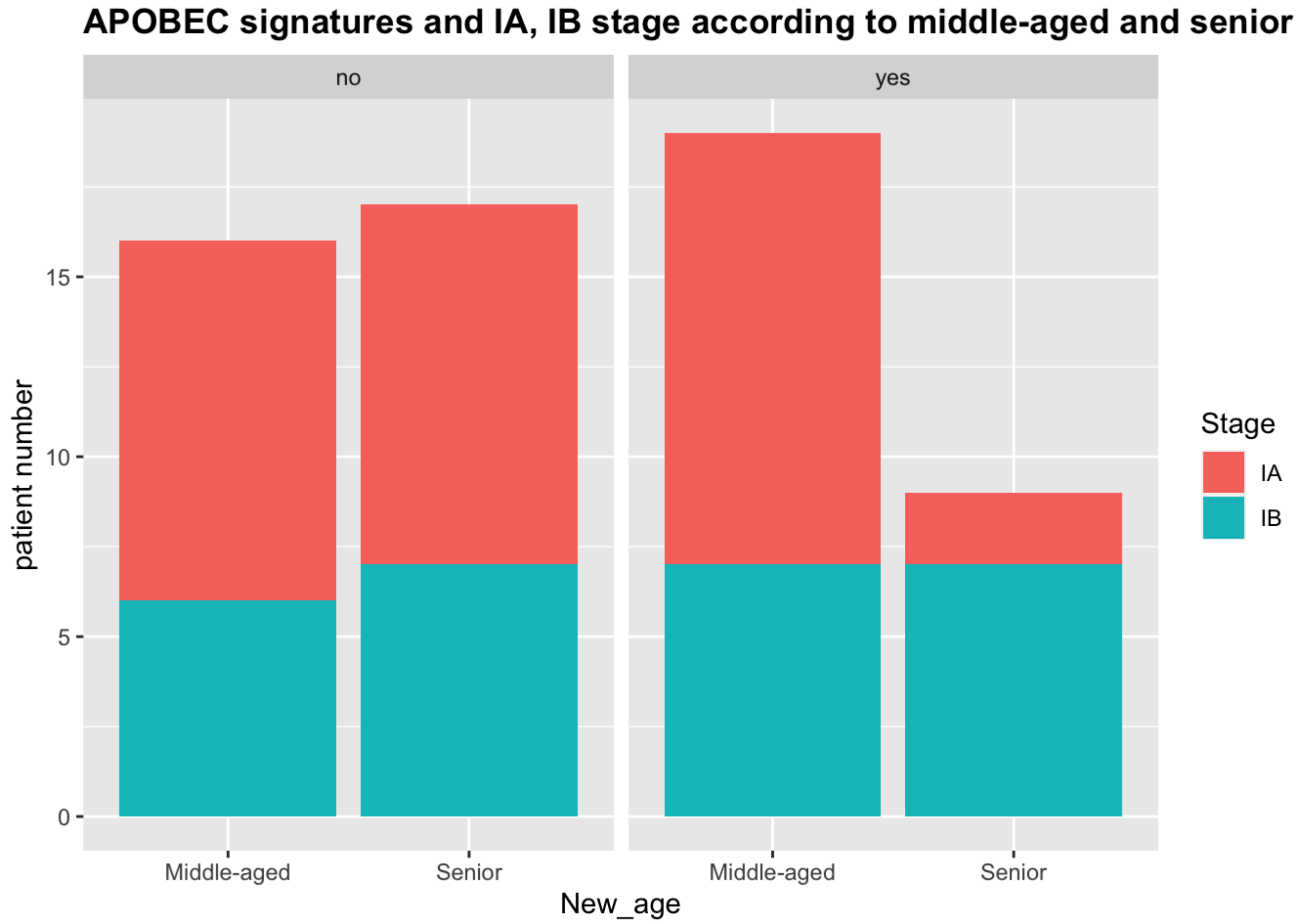
APOBEC_Enriched	OverallStaging	Ages	n
no	>=III	50s	1
no	>=III	60s	1
no	>=III	70s	3
no	IA	40s	1
no	IA	50s	5
no	IA	60s	7

```
cc <- d6_2_data %>% group_by(APOBEC_Enriched) %>% count(OverallStaging, Ages) %>%
  ggplot(aes(Ages, n, fill=OverallStaging)) +
  geom_bar(stat="identity", position="stack") +
  facet_grid(~APOBEC_Enriched) +
  theme(legend.position = "none") +
  labs(title = '') +
  ylab("patient number")

dd <- d6_2_data %>% group_by(APOBEC_Enriched) %>% count(OverallStaging, Ages) %>%
  ggplot(aes(Ages, n, fill=OverallStaging)) +
  geom_bar(stat="identity", position="fill") +
  facet_grid(~APOBEC_Enriched) +
  ylab("patient rate") +
  labs(title = 'APOBEC signatures and stage in nonsmokers') +
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 5.75)) +
  theme(legend.direction = 'horizontal',
        legend.position = c(0.5, 1.11),
        legend.margin = margin(0, 0, 0, 0),
        legend.key.size = unit(3, 'mm'),
        legend.title = element_text(size=9))

grid.arrange(cc, dd, ncol = 2)
```



A new data classification criterion is prepared to identify more distinguishable features. After filtering only IA, IB, IIA, and IIB among tumor stages, IA and IIA are classified as IA, and IB and IIB are classified as IB. As for the age of the patient, those over 65 years of age, the international standard for the elderly, are classified as Senior and those under 65 are classified as Middle-aged.

```
d6_2_data %>% filter(SmokingStatus=="no") %>% filter(OverallStaging %in% c("IA", "IB"
, "IIA", "IIB")) %>% mutate(Stage=case_when(OverallStaging %in% c("IA", "IIA") ~ "IA"
, OverallStaging %in% c("IB", "IIB") ~ "IB")) %>% mutate(New_age=ifelse(Age<65, "Midd
le-aged", "Senior")) %>% group_by(APOBEC_Enriched) %>% count(Stage, New_age) %>%
  ggplot(aes(New_age, n, fill=Stage)) +
  geom_bar(stat="identity") +
  facet_grid(~APOBEC_Enriched) +
  ylab("patient number") +
  labs(title = 'APOBEC signatures and IA, IB stage according to middle-aged and senio
r') +
  theme(plot.title = element_text(face = "bold"))
```



“yes” means “APOBEC_Enriched”, and “no” means “not APOBEC_Enriched”.

#4.3. Result and analysis We can see that there is only a slight difference in the stage of the tumor depending on the abundance of APOBEC. The APOBEC-enriched side had slightly fewer cases than the non-APOBEC-enriched group. Tumors of stage IA were more common in those not rich in APOBE, and other tumor stages showed similar patterns regardless of APOBEC. There was a difference in tumor stage according to the abundance of APOBEC according to age. In the APOBEC-rich group, the incidence rate was relatively high in those in their 40s, and the incidence was relatively low in those in their 60s and 70s. Looking at the last plot in 4.2, in the case of APOBEC enrichment, it can be inferred that IA stage tumors are more lethal than IB stage tumors in elderly patients. If we look at the case where tumor stage and age are divided by only two criteria, the number of middle-aged and elderly patients is similar in the group without APOBEC enrichment. In the APOBEC enrich group, the number of IA stage patients was significantly less in the elderly group than in the middle-aged group.

#5. Conclusion and Suggestion Only weak correlation was found between APOBEC signatures and tumor stage, but no strong correlation was found. Therefore, it is difficult to regard it as a meaningful relationship. As a result, The correlation between APOBEC enrichment and tumor stage is difficult to recognize, and it was found that the probability of developing IA stage tumor in the APOBEC enrichment situation was slightly higher than in the non-enrichment situation. In the case of APOBEC enrichment, it can be inferred that IA stage tumors are more lethal than IB stage tumors in elderly patients. It is hoped that the correlation between APOBEC and lung cancer can be closely clarified by using criteria that vary according to the degree of enrichment of APOBEC. In addition, it is hoped that it can be visualized using factors related to the tumor stage, the patient’s age, and whether or not to smoke.