

Survival Analysis- Celiac

Project1: Survival analysis-Celiac

Description: Celiac patients with Colorectal Cancer (CRC)

In this project we investigate if patients with celiac disease are at a higher risk of future colorectal cancer (CRC) compared to controls. The start of CRC is defined as the day of the first ICD (International Classification of Diseases) code starting by *C18*, *C19*, or *C20*. End of follow-up of the study is *2019-12-31*. The study design is a **matched cohort study**, where each celiac patient was matched to maximum 2 non-celiac persons on *age* and *gender* at start of follow-up.

Definitions

The International Classification of Diseases (ICD) codes used to identify colorectal cancer and celiac disease in patients are as follows:

- **ICD-10 Codes:**
- **C18:** Malignant neoplasm of colon
- **C19:** Malignant neoplasm of rectosigmoid junction
- **C20:** Malignant neoplasm of rectum
- **C21:** Malignant neoplasm of anus and anal canal

Data

The data comes in the following three files:

1. celiac.txt
2. nonceliac.txt
3. ICD.txt

Variables

The file named ***celiac.txt*** consists of even variables as follows:

- *Id.number*: The number identifying the celiac patient
- *Start.of.flwp*: Start of follow-up
- *Gender*: 1= Male, 2= Female
- *Age*: The person's age
- *Birthday*: The person's birthday
- *Date.of.Death*: The day the person died
- *Date.of.Migration*: The day the person migrated

The file named ***nonceliac.txt*** contains the following four variables:

- *Id.number*: The number identifying the non-celiac person
- *Celiac.number*: The number identifying the celiac patient the person is matched to
- *Date.of.Death*: The day the person died
- *Date.of.Migration*: The day the person migrated.

The file named ***ICD.txt*** contains the following three variables:

- *Id.number*: The number identifying the person
- *ICD.codes*: The list of ICD codes assigned at a specific date.
- *Date.of.ICD.codes*: The date the ICD codes were assigned

Tasks

Estimate the following measures both for cases, i.e. celiac patients and controls, i.e., non-celiac patients

1. *Number of persons* per group (N)
2. *Number of person years* per group (PY)
3. *Number of events* per group (Events)
4. *Incidence rate* per 1000 person-year 95% confidence intervals

5. The *Hazard ratios* (HR) with 95% confidence intervals
 - a. without adjusting for *Age* and *Gender* variables
 - b. adjusting for *Age* and *Gender* variables
6. Consolidate the above results into a table

Data wrangling and cleaning

```
library(survival)
#library(survminer)
library(simstudy) #the package for non-PH models
library(ggplot2)
library(dplyr)
library(tidyverse)
library(viridis)
library(ggrepel)
library(ggthemes)
library(lubridate) # to handle the dates data
library(anytime) #handles the dates data in any format
library(fmsb) # for 95% CI of incidence rate
library(gt)
```

1. Upload the data files

```
celiac <- read.csv("~/Desktop/Celiac-main/celiac.txt")

nonceliac <- read.csv("~/Desktop/Celiac-main/nonceliac.txt")

ICD <- read.csv("~/Desktop/Celiac-main/ICD.txt")
# peek into the data
glimpse(celiac)
```

Rows: 12,350

Columns: 7

```
$ id.number      <int> 424, 881, 1739, 2462, 2652, 2785, 3136, 3192, 3888, ~
$ Start.of.flwp  <int> 20081106, 20070613, 20040609, 20000306, 20040712, 20~
$ Gender         <int> 2, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, 2~
$ Age           <int> 72, 21, 12, 19, 46, 17, 24, 31, 27, 17, 38, 7, 77, 7~
$ Birthday       <chr> "1936-05-15", "1986-02-15", "1991-12-15", "1980-09-1~
$ Date.of.Death  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

```
$ Date.of.Migration <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

```
glimpse(nonceliac)
```

```
Rows: 24,605
```

```
Columns: 4
```

```
$ id.number          <int> 7659518, 1038521, 9265189, 10658482, 8460019, 109809~
```

```
$ Celiac.number      <int> 424, 424, 881, 881, 1739, 1739, 2462, 2462, 2652, 26~
```

```
$ Date.of.Death      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20081016~
```

```
$ Date.of.Migration <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

```
glimpse(ICD)
```

```
Rows: 47,407
```

```
Columns: 3
```

```
$ id.number          <int> 2855591, 2855591, 5552533, 9515567, 2112819, 9075074~
```

```
$ ICD.codes          <chr> "C402", "C402", "C787 C809 M069", "D489 C921", "Z489~
```

```
$ Date.of.ICD.codes <chr> "2003-03-04", "2002-12-25", "2010-05-25", "2001-09-2~
```

2. Dates columns need to be in one format

```
celiac <- celiac %>%  
  mutate(across(c(5:7), as.character))%>%  
  mutate(across(c(2,5,6,7), anydate))  
  
nonceliac <- nonceliac %>%  
  mutate(across(c(3,4), as.character))%>%  
  mutate(across(c(3,4), anydate))  
  
ICD <-ICD%>%  
  mutate(Date.of.ICD.codes= as.Date(Date.of.ICD.codes))  
  
#check the data sets  
glimpse(celiac)
```

```
Rows: 12,350
```

```
Columns: 7
```

```
$ id.number          <int> 424, 881, 1739, 2462, 2652, 2785, 3136, 3192, 3888, ~
```

```
$ Start.of.flwp      <date> 2008-11-06, 2007-06-13, 2004-06-09, 2000-03-06, 200~
```

```
$ Gender             <int> 2, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, 2~
```

```
$ Age          <int> 72, 21, 12, 19, 46, 17, 24, 31, 27, 17, 38, 7, 77, 7~
$ Birthday     <date> 1936-05-15, 1986-02-15, 1991-12-15, 1980-09-15, 195~
$ Date.of.Death <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Date.of.Migration <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
glimpse(nonceliac)
```

```
Rows: 24,605
Columns: 4
$ id.number      <int> 7659518, 1038521, 9265189, 10658482, 8460019, 109809~
$ Celiac.number  <int> 424, 424, 881, 881, 1739, 1739, 2462, 2462, 2652, 26~
$ Date.of.Death  <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2008-10~
$ Date.of.Migration <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
glimpse(ICD)
```

```
Rows: 47,407
Columns: 3
$ id.number      <int> 2855591, 2855591, 5552533, 9515567, 2112819, 9075074~
$ ICD.codes      <chr> "C402", "C402", "C787 C809 M069", "D489 C921", "Z489~
$ Date.of.ICD.codes <date> 2003-03-04, 2002-12-25, 2010-05-25, 2001-09-20, 200~
```

3. Extract CRC events data from the ICD file

```
# Find out the CRC events from ICD file both for celiac and nonceliac
ICD_CRC_events <- ICD %>%
  filter(grepl('C18', ICD.codes))%>%
  rbind(
    ICD %>% filter(grepl('C19', ICD.codes)) %>%
    rbind(
      ICD %>% filter(grepl('C20', ICD.codes)) %>%
      mutate(Date.of.ICD.codes= as.Date(Date.of.ICD.codes))
    )
  )

head(ICD_CRC_events)
```

	id.number	ICD.codes	Date.of.ICD.codes
1	10447614	C187 C787 C771 C722 I109 E119	2009-04-20
2	3912733	C187	2005-01-24
3	2040543	C189	2005-01-28
4	4780187	C186 I109 E785 K900A	2010-08-25
5	344621	C186 I489B I209 N390	2008-06-07
6	7462794	Z514 C183	2010-01-06

4. Construct the data frame for celiac patients by following the instructions given in the project description

```
#Find first CRC events both for celiac and nonceliac groups
celiac_CRC <- celiac%>%
  filter(id.number %in% unique(ICD_CRC_events$id.number) ) %>%
  merge(ICD_CRC_events%>%
    filter(id.number %in% unique(celiac$id.number))%>%
    group_by(id.number)%>%
    summarise(Date.of.CRC.onset = min(Date.of.ICD.codes) ),
    by.x = c("id.number"),
    by.y = c("id.number"),
    all.x = TRUE)%>%
  filter(Start.of.flwp <= Date.of.CRC.onset)%>%
  # exclude the CRC cases before the follow-up
  mutate(Date.of.flwp.end = do.call(pmin, c(select(., c(6:8)),
    na.rm = TRUE)) ) %>%

  mutate(person.time.years =
    as.numeric(Date.of.CRC.onset-Start.of.flwp)/ 365.25)%>%
  # calculate person-years time
  mutate(CRC.status= "Yes")
#mutate the variable for the CRC events

celiac_nonCRC <- celiac %>%
  filter(id.number %in% setdiff(unique(celiac$id.number),unique(ICD_CRC_events$id.number) ))%>%
  mutate(Date.of.CRC.onset= as.Date(NA))%>%
  mutate(Date.of.flwp.end = as.Date("2019-12-31") ) %>%
  #filter(Start.of.flwp <= Date.of.CRC.onset)%>%
  mutate(Date.of.flwp.end = do.call(pmin, c(select(., c(6:9)),
    na.rm = TRUE)) ) %>%

  #filter(Start.of.flwp>Date.of.flwp.end)%>%
  mutate(person.time.years =
    as.numeric(Date.of.flwp.end-Start.of.flwp)/ 365.25)%>%
  mutate(CRC.status= "No")
#mutate the variable for the CRC events

celiac_final_df <- rbind(celiac_CRC, celiac_nonCRC)%>%
  filter(Start.of.flwp<Date.of.flwp.end)%>%
  # filter out the events with negative follow-up time
  mutate(CRC.status= as.factor(CRC.status))%>%
  mutate(exposure.status= "1")
#mutate the variable for the exposure status (celiac~1, nonceliac~0)
```

```
dim(celiac_CRC)
```

```
[1] 113  11
```

```
dim(celiac_nonCRC)
```

```
[1] 12199  11
```

```
dim(celiac_final_df)
```

```
[1] 12310  12
```

Modelling and analysis

5. Calculate the required estimates (*events*, *person years time/1000* and *incidence rates*) for celiac patients

```
total_cases_celiac = length(celiac_CRC$id.number)
person_time_celiac = sum(celiac_final_df$person.time.years)
Incidence_rate_celiac = total_cases_celiac/ person_time_celiac *1000

total_cases_celiac
```

```
[1] 113
```

```
person_time_celiac
```

```
[1] 137267.8
```

```
## 95% CI for incidence rate using fmsb package
IRCI(total_cases_celiac*1000, person_time_celiac, conf.level=0.95)
```

```
$IR  
[1] 0.8232082
```

```
$IRL  
[1] 0.8184085
```

```
$IRU  
[1] 0.828008
```

```
# OR using the following lines  
SE_celiac = 1.96 * sqrt((total_cases_celiac*1000)/person_time_celiac^2 )  
CIL_95_celiac_IR <- Incidence_rate_celiac-SE_celiac  
CIU_95_celiac_IR <- Incidence_rate_celiac+SE_celiac
```

6. Construct the data frame for non-celiac patients based on the instructions

```
nonceliac_CRC <-nonceliac %>%  
  merge(celiac[,c(1:5)],  
        by.x = "Celiac.number" ,  
        by.y = "id.number",  
        all.x = TRUE )%>%  
  filter(id.number %in% unique(ICD_CRC_events$id.number) ) %>%  
  merge( ICD_CRC_events)%>%  
    filter(id.number %in% unique(nonceliac$id.number)) %>%  
  group_by(id.number)%>%  
    summarise(Date.of.CRC.onset = min(Date.of.ICD.codes)),  
  by.x = c("id.number"),  
  by.y =c("id.number"),  
  all.y = TRUE  
)%>%  
select(c(1,5:8,3,4,9))%>%  
filter(Start.of.flwp <= Date.of.CRC.onset)%>%  
mutate(  
  Date.of.flwp.end =  
    do.call(pmin, c(select(., c(6:8)),  
                    na.rm = TRUE))  
)%>%  
filter(Start.of.flwp<Date.of.flwp.end)%>%  
mutate(person.time.years = as.numeric(Date.of.CRC.onset-Start.of.flwp)/ 365.25)%>%  
mutate(CRC.status= "Yes")  
  
nonceliac_nonCRC <- nonceliac%>%
```



```

#match the nonceliac patients with celiac with celiac number
merge(celiac[,c(1:5)],
      by.x = "Celiac.number" ,
      by.y = "id.number",
      all.x = TRUE )%>%
#filter the nonceliacs found in the ICD data
filter(id.number %in%
       setdiff(unique(nonceliac$id.number),
               unique(ICD_CRC_events$id.number) )
       )%>%
#mutate 'Date.of.CRC.onset= NA' variable
mutate(Date.of.CRC.onset= as.Date(NA))%>%
#mutate Followup end date variable with fixed date(2019-12-31)
mutate(Date.of.flwp.end = as.Date("2019-12-31") )%>%
select(c(2,5:8,3:4,9,10))%>%
mutate(Date.of.flwp.end =
       do.call(pmin, c(select(., c(6:9)),
                       na.rm = TRUE))
       )%>%
filter( Start.of.flwp < Date.of.flwp.end)%>%
mutate( person.time.years =
       as.numeric(Date.of.flwp.end-Start.of.flwp)
       / 365.25)%>%
mutate(CRC.status= "No")

nonceliac_final_df <- rbind(nonceliac_CRC, nonceliac_nonCRC)%>%
  filter(Start.of.flwp<Date.of.flwp.end)%>%
  mutate(CRC.status= as.factor(CRC.status))%>%
  mutate(exposure.status= "0")

```

7. Calculate the required estimates (*events*, *person years time /1000* and *incidence rate*) for nonceliac controls

```

total_cases_nonceliac = length(nonceliac_CRC$id.number)
person_time_nonceliac = sum(nonceliac_final_df$person.time.years)
Incidence_rate_nonceliac = (total_cases_nonceliac/ person_time_nonceliac *1000)

total_cases_nonceliac

```

[1] 194

```
person_time_nonceliac
```

```
[1] 272313.7
```

```
## 95% CI for incidence rate using fmsb package  
IRCI(total_cases_nonceliac*1000, person_time_nonceliac, conf.level=0.95)
```

```
$IR
```

```
[1] 0.7124136
```

```
$IRL
```

```
[1] 0.7092435
```

```
$IRU
```

```
[1] 0.7155838
```

```
# OR
```

```
SE_nonceliac = 1.96*sqrt(  
  (total_cases_nonceliac*1000)/person_time_nonceliac^2 )  
CIL_95_nonceliac_IR <- Incidence_rate_nonceliac-SE_nonceliac  
CIL_95_nonceliac_IR <- Incidence_rate_nonceliac+SE_nonceliac
```

Construct the final data frame to estimate the *hazard ratios* (HR) using *Cox proportional hazard* (Cox-PH) model

```
df_celiac_nonceliac <- rbind(celiac_final_df,nonceliac_final_df)%>%  
mutate(flwp.time = as.numeric((Date.of.flwp.end-Start.of.flwp))  
  )%>%  
  mutate(exposure.status=as.factor(exposure.status))%>%  
  mutate(Gender= as.factor(Gender))%>%  
  filter(Start.of.flwp < Date.of.flwp.end)%>%  
  select(c(1:10,12,11,13))  
  
head(df_celiac_nonceliac)
```

	id.number	Start.of.flwp	Gender	Age	Birthday	Date.of.Death	Date.of.Migration
1	18671	2002-02-01	1	77	1924-11-15	2003-08-27	<NA>
2	41943	2000-03-22	2	74	1925-08-15	2004-04-15	<NA>
3	170391	2000-05-03	1	37	1963-01-15	2017-07-09	<NA>
4	183428	2005-06-01	2	34	1970-11-15	<NA>	<NA>

5	199935	2005-03-23	1	68	1937-02-15	2006-01-01	<NA>
6	224099	2010-06-09	1	76	1933-08-15	2015-02-03	<NA>

	Date.of.CRC.onset	Date.of.flwp.end	person.time.years	exposure.status
1	2002-04-09	2002-04-09	0.1834360	1
2	2003-04-21	2003-04-21	3.0800821	1
3	2015-08-31	2015-08-31	15.3264887	1
4	2014-04-29	2014-04-29	8.9089665	1
5	2005-08-21	2005-08-21	0.4134155	1
6	2013-01-21	2013-01-21	2.6201232	1

	CRC.status	flwp.time
1	Yes	67
2	Yes	1125
3	Yes	5598
4	Yes	3254
5	Yes	151
6	Yes	957

Survival analysis- Cox model

Apply *Cox-PH model* -unadjusted for *Age* and *Gender* variables

```
HR_CRC_Unadjusted <- coxph(Surv( as.numeric(Start.of.flwp),
                                as.numeric(Date.of.flwp.end),
                                as.numeric(CRC.status))
~exposure.status, data=df_celiac_nonceliac)

summary(HR_CRC_Unadjusted)
```

Call:

```
coxph(formula = Surv(as.numeric(Start.of.flwp), as.numeric(Date.of.flwp.end),
                    as.numeric(CRC.status)) ~ exposure.status, data = df_celiac_nonceliac)
```

n= 36668, number of events= 307

	coef	exp(coef)	se(coef)	z	Pr(> z)
exposure.status1	0.1443	1.1552	0.1183	1.219	0.223

	exp(coef)	exp(-coef)	lower .95	upper .95
exposure.status1	1.155	0.8657	0.916	1.457

Concordance= 0.497 (se = 0.014)

```

Likelihood ratio test= 1.47 on 1 df, p=0.2
Wald test              = 1.49 on 1 df, p=0.2
Score (logrank) test = 1.49 on 1 df, p=0.2

```

```

#Test PH assumption
test.ph <- cox.zph(HR_CRC_Unadjusted)

test.ph

```

```

              chisq df      p
exposure.status 16.3  1 5.4e-05
GLOBAL          16.3  1 5.4e-05

```

```

#ggcoxzph(test.ph)

```

Next, apply the *Cox-PH model* adjusted for *Age* and *Gender* variables

```

HR_CRC_AgeGender_adjusted <- coxph(
  Surv(flpw.time,as.numeric(CRC.status)) ~exposure.status+Age+Gender,
                                data=df_celiac_nonceliac
                                )

summary(HR_CRC_AgeGender_adjusted)

```

Call:

```

coxph(formula = Surv(flpw.time, as.numeric(CRC.status)) ~ exposure.status +
      Age + Gender, data = df_celiac_nonceliac)

```

n= 36668, number of events= 307

```

              coef exp(coef)  se(coef)      z Pr(>|z|)
exposure.status1  0.135881  1.145546  0.118374  1.148    0.251
Age               0.072753  1.075465  0.003578 20.332 <2e-16 ***
Gender2          -0.154423  0.856910  0.114902 -1.344    0.179
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

              exp(coef) exp(-coef) lower .95 upper .95
exposure.status1      1.1455      0.8729      0.9083      1.445
Age                   1.0755      0.9298      1.0679      1.083

```

Gender2	0.8569	1.1670	0.6841	1.073
---------	--------	--------	--------	-------

Concordance= 0.888 (se = 0.006)

Likelihood ratio test= 668.5 on 3 df, p=<2e-16

Wald test = 420.9 on 3 df, p=<2e-16

Score (logrank) test = 665.8 on 3 df, p=<2e-16

```
PH_assump_CRC_AgeGender_Adjusted <- cox.zph(HR_CRC_AgeGender_adjusted)
PH_assump_CRC_AgeGender_Adjusted
```

	chisq	df	p
exposure.status	21.949	1	2.8e-06
Age	2.814	1	0.093
Gender	0.489	1	0.484
GLOBAL	24.748	3	1.7e-05

```
 #(PH_assump_CRC_AgeGender_Adjusted)
```

Stratify the followup duration time

```
#Divide the followup duration into 3 intervals [0,200), [200, 1650), [1650, 7302)
```

```
#Here, we use 'survSplit' function from 'simstudy' package
```

```
df_celiac_nonceliac_stratfied <-
```

```
  survSplit(Surv(flpw.time, as.numeric(CRC.status)) ~.,
            data= df_celiac_nonceliac, cut=c(200,1650),
            episode= "tgroup")
```

```
head(df_celiac_nonceliac_stratfied)
```

	id.number	Start.of.flwp	Gender	Age	Birthday	Date.of.Death	Date.of.Migration
1	18671	2002-02-01	1	77	1924-11-15	2003-08-27	<NA>
2	41943	2000-03-22	2	74	1925-08-15	2004-04-15	<NA>
3	41943	2000-03-22	2	74	1925-08-15	2004-04-15	<NA>
4	170391	2000-05-03	1	37	1963-01-15	2017-07-09	<NA>
5	170391	2000-05-03	1	37	1963-01-15	2017-07-09	<NA>
6	170391	2000-05-03	1	37	1963-01-15	2017-07-09	<NA>
	Date.of.CRC.onset	Date.of.flwp.end	person.time.years	exposure.status	tstart		
1	2002-04-09	2002-04-09	0.183436	1	0		
2	2003-04-21	2003-04-21	3.080082	1	0		

3	2003-04-21	2003-04-21	3.080082	1	200
4	2015-08-31	2015-08-31	15.326489	1	0
5	2015-08-31	2015-08-31	15.326489	1	200
6	2015-08-31	2015-08-31	15.326489	1	1650

	flwp.time	event	tgroup
1	67	1	1
2	200	0	1
3	1125	1	2
4	200	0	1
5	1650	0	2
6	5598	1	3

Apply again the *Cox-PH model* on *stratified* followup duration- unadjusted for *Age* and *Gender*

```
HR_CRC_Unadjusted_stratfied <- survival::coxph(Surv(tstart,flwp.time, event)~
  exposure.status*strata(tgroup),
  data=df_celiac_nonceliac_stratfied)

summary(HR_CRC_Unadjusted_stratfied)
```

Call:

```
survival::coxph(formula = Surv(tstart, flwp.time, event) ~ exposure.status *
  strata(tgroup), data = df_celiac_nonceliac_stratfied)
```

n= 106558, number of events= 307

	coef	exp(coef)	se(coef)	z
exposure.status1	2.69983	14.87716	0.53229	5.072
exposure.status1:strata(tgroup)tgroup=2	-2.50733	0.08149	0.56974	-4.401
exposure.status1:strata(tgroup)tgroup=3	-3.06894	0.04647	0.56035	-5.477
	Pr(> z)			
exposure.status1	3.93e-07	***		
exposure.status1:strata(tgroup)tgroup=2	1.08e-05	***		
exposure.status1:strata(tgroup)tgroup=3	4.33e-08	***		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower	.95
exposure.status1	14.87716	0.06722	5.24123	
exposure.status1:strata(tgroup)tgroup=2	0.08149	12.27207	0.02668	

```

exposure.status1:strata(tgroup)tgroup=3    0.04647    21.51909    0.01550
                                         upper .95
exposure.status1                          42.2286
exposure.status1:strata(tgroup)tgroup=2    0.2489
exposure.status1:strata(tgroup)tgroup=3    0.1394

```

```

Concordance= 0.563 (se = 0.015 )
Likelihood ratio test= 49.8 on 3 df, p=9e-11
Wald test              = 31.07 on 3 df, p=8e-07
Score (logrank) test = 51.08 on 3 df, p=5e-11

```

```
cox.zph(HR_CRC_Unadjusted_stratfied)
```

```

                                chisq df    p
exposure.status                0.616  1 0.43
exposure.status:strata(tgroup) 0.570  2 0.75
GLOBAL                        5.712  3 0.13

```

```
#ggcoxzph(HR_CRC_Unadjusted_stratfied)
```

Apply again the *Cox-PH model* on stratified followup duration- adjusted for *Age* and *Gender* variables

```

HR_CRC_AgeGender_adjusted_stratified <- survival::coxph(Surv(tstart,flwp.time, event)~
  exposure.status*strata(tgroup)+Age+Gender,
  data=df_celiac_nonceliac_stratfied)

summary(HR_CRC_AgeGender_adjusted_stratified)

```

Call:

```
survival::coxph(formula = Surv(tstart, flwp.time, event) ~ exposure.status *
  strata(tgroup) + Age + Gender, data = df_celiac_nonceliac_stratfied)
```

n= 106558, number of events= 307

```

                                coef exp(coef) se(coef)      z
exposure.status1                2.664553 14.361534  0.532301   5.006
Age                             0.072778  1.075492  0.003593  20.254
Gender2                         -0.153608  0.857608  0.114911  -1.337
exposure.status1:strata(tgroup)tgroup=2 -2.480087  0.083736  0.569740  -4.353

```

```

exposure.status1:strata(tgroup)tgroup=3 -3.037078  0.047975  0.560359 -5.420
                                         Pr(>|z|)
exposure.status1                        5.57e-07 ***
Age                                     < 2e-16 ***
Gender2                                0.181
exposure.status1:strata(tgroup)tgroup=2 1.34e-05 ***
exposure.status1:strata(tgroup)tgroup=3 5.96e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

	exp(coef)	exp(-coef)	lower .95
exposure.status1	14.36153	0.06963	5.05947
Age	1.07549	0.92981	1.06794
Gender2	0.85761	1.16603	0.68466
exposure.status1:strata(tgroup)tgroup=2	0.08374	11.94230	0.02741
exposure.status1:strata(tgroup)tgroup=3	0.04797	20.84425	0.01600

	upper .95
exposure.status1	40.7658
Age	1.0831
Gender2	1.0742
exposure.status1:strata(tgroup)tgroup=2	0.2558
exposure.status1:strata(tgroup)tgroup=3	0.1439

```

Concordance= 0.896 (se = 0.006 )
Likelihood ratio test= 715.7 on 5 df,  p=<2e-16
Wald test               = 445.5 on 5 df,  p=<2e-16
Score (logrank) test = 714.5 on 5 df,  p=<2e-16

```

```
cox.zph(HR_CRC_AgeGender_adjusted_stratified)
```

	chisq	df	p
exposure.status	0.623	1	0.43
Age	2.464	1	0.12
Gender	0.453	1	0.50
exposure.status:strata(tgroup)	0.570	2	0.75
GLOBAL	8.436	5	0.13

```
#ggcoxzph(HR_CRC_AgeGender_adjusted_stratified)
```

Table

Now, construct the final table of the results

Table 1 combines the completed tasks of the project.

```
# Data preparation

summary_table <- tibble(
  Group = c("Celiac Patients", "Non-Celiac Controls"),
  `Number of Persons` = c(nrow(celiac_final_df), nrow(nonceliac_final_df)),
  `Number of Person-Years` = c(person_time_celiac, person_time_nonceliac),
  `Number of Events` = c(total_cases_celiac, total_cases_nonceliac),
  `Incidence Rate per 1000 PY (95% CI)` = c(
    paste0(round(Incidence_rate_celiac, 2), " (", round(CIL_95_celiac_IR, 2), " - ", round(CIL_95_nonceliac_IR, 2), " (", round(Incidence_rate_nonceliac, 2), " (", round(CIL_95_nonceliac_IR, 2), " - ", round(CIL_95_celiac_IR, 2), " )"),
    ),
  `Unadjusted HR (95% CI)` = c(
    paste0(round(summary(HR_CRC_Unadjusted)$coefficients[1, "exp(coef)"], 2), " (", round(summary(HR_CRC_Unadjusted)$coefficients[2, "exp(coef)"], 2), " )"),
    ),
  `Adjusted HR (95% CI)` = c(
    paste0(round(summary(HR_CRC_AgeGender_adjusted)$coefficients[1, "exp(coef)"], 2), " (", round(summary(HR_CRC_AgeGender_adjusted)$coefficients[2, "exp(coef)"], 2), " )"),
    )
)

# Create the table using gt
gt_table <- summary_table %>%
  gt() %>%
  tab_header(
    title = "Summary of Colorectal Cancer Risk among Celiac and Non-Celiac Patients",
    subtitle = "Based on Survival Analysis"
  ) %>%
  cols_label(
    `Number of Persons` = "N",
    `Number of Person-Years` = "PY",
    `Number of Events` = "Events",
    `Incidence Rate per 1000 PY (95% CI)` = "Incidence Rate / 1000 PY (95% CI)",
    `Unadjusted HR (95% CI)` = "Unadjusted HR (95% CI)",
    `Adjusted HR (95% CI)` = "Adjusted HR (95% CI)"
  ) %>%
  fmt_number(
    columns = c(`Number of Person-Years`, `Number of Events`, `Unadjusted HR (95% CI)`, `Adjusted HR (95% CI)`),
    decimals = 2
  ) %>%
  cols_align(
```

```

    align = "center",
    columns = everything()
)

# Display the table
gt_table

```

Table 1: Results

Summary of Colorectal Cancer Risk among Celiac and Non-Celiac Patients
Based on Survival Analysis

Group	N	PY	Events	Incidence Rate / 1000 PY (95% CI)	Unadjusted
Celiac Patients	12310	137,267.82	113.00	0.82 (0.82 - 0.83)	1.16 (0.82 - 1.64)
Non-Celiac Controls	24358	272,313.71	194.00	0.71 (0.72 - 0.72)	0.82 (0.72 - 0.94)

Interpretations and conclusions

Overall, we observed an increased risk of CRC among celiac patients compared to the nonceliac individuals which was statistically non-significant. The unadjusted and adjusted (for age and sex) hazards were approximately *15%* and *16%* higher among celiac patients compared to the non-celiac individuals (Table 1). However, the proportional-hazard (PH) assumptions for the cox regression model were not fulfilled ($p < 0.05$) when tested with Schoenfeld test. To tackle this, we stratified the follow-up duration into three intervals. Subsequently, the CRC hazards in exposed group increased significantly ($p < 0.05$) up to 15-times (Table 1) compared to the unexposed group for the first time-interval. In contrast, the CRC hazard in celiac patients was significantly lower ($p < 0.05$) compared to the unexposed for other time-intervals. Similar findings were observed after adjusting for age and sex (Table 1). The results should be interpreted carefully due to large uncertainties in the estimates.