Clinical Trial Statistical Data Analysis Code

Jason Massey

2025-02-24

Libraries and Reading in Data

```
####################################
#    Downloading Libraries         #
####################################
library(tidyverse)
library(summarytools)
library(dplyr)
library(eeptools)
library(table1)

####################################
#    Reading in Datasets           #
####################################
demographics <- read.csv("XXXX/Patient_Demographics.csv")
diagnosis <- read.csv("XXXX/Patient_Diagnosis.csv")
treatment <- read.csv("XXXX/Patient_Treatment.csv")
```

Question 1

First, the clinic would like to know the number and proportion of patients diagnosed with each type of cancer (i.e., breast cancer vs colon cancer). Note that patients can be diagnosed with more than one type of cancer; these patients should be categorized as having "both breast and colon cancer".

```
# Look at the Data
dfSummary(diagnosis, style = "grid", plain.ascii = TRUE)

# Create dataframe removing duplicate cancer diagnosis by types per patient
diagnosis.type <- diagnosis [!duplicated(diagnosis[c(1,4)]),]

# Label Cancer Type as Colon, Breast, or Both
diagnosis.type$type <- ifelse(duplicated(diagnosis.type$patient_id) |
duplicated(diagnosis.type$patient_id,fromLast=TRUE),
                              "both", diagnosis.type$diagnosis )

# Subset to remaining non-duplicated patients dataframe
diagnosis.type2 <- diagnosis.type %>% distinct(patient_id, .keep_all = TRUE)

# Look at counts of cancer diagnosis by type
dfSummary(diagnosis.type2, style = "grid", plain.ascii = TRUE)
```

Solution N(%): Breast cancer = 31 (66.0%)
Colon Cancer = 11 (23.4%)
Both = 5 (10.6%)

Question 2

The clinic wants to know how long it takes for patients to start therapy after being diagnosed, which they consider to be helpful in understanding the quality of care for the patient. How long after their earliest diagnosis do patients start treatment? Please provide the mean, median, minimum, and maximum for each of the cancer groups identified in Question 1.

```
# Group by patient ID, order by date, and filtering by row per first diagnosis date
diagnosis.date <- diagnosis %>%
  group_by(patient_id) %>%
  arrange(diagnosis_date) %>%
 filter(row_number()==1)

# Group by patient ID, order by date, and filtering by row per first treatment date
treatment.date <- treatment %>%
  group_by(patient_id) %>%
  arrange(treatment_date) %>%
filter(row_number()==1)
```

```
# Join 1st diagnosis date, 1st treatment date, and cancer types by patient_ID
# NOTE: It appears 1 patient didn't receive treatment
days <- diagnosis.type2 %>%
  left_join(diagnosis.date, by='patient_id') %>%
  left_join(treatment.date, by='patient_id')

# Calculate time to treatment: time = (treatment date - diagnosis date)
days$days <- as.numeric(as.Date(days$treatment_date, "%m/%d/%y"))-
  as.numeric(as.Date(days$diagnosis_date.x, "%m/%d/%y"))

# Display summary statistics of days grouping by cancer type
tapply(days$days, days$type, summary)
```

Solution (days to treatment):

```
           Both
Min.  Median  Mean   Max.
14.0  49.0    51.6   91.0
        Breast Cancer
Min.  Median  Mean   Max.
0.0   8.00    37.70  153.00
        Colon Cancer
Min.  Median  Mean   Max.
0.00  14.00   80.18  366.00
```

Question 3

A patient’s first-line treatment is the drug (i.e., monotherapy) or set of drugs (i.e., combination therapy) that the patient received at the start of systemic treatment`` (e.g.,first treatment after earliest diagnosis) for their disease (for more information on first-line treatments, click here). Without access to information about the clinician’s specific decision making, we can infer a patient’s first-line treatment regimen based on the drug or set of drugs they received in their first treatment instance. Using this approach, which treatment regimens [i.e., drug(s)] do you think would be indicated as first-line treatment for patients with...

○ breast cancer only?
○ colon cancer only?
○ both breast and colon cancer?

```
# Inner join days to the original treatment dataset to consider combination therapy
firstline <- days %>%
  inner_join(treatment, by= c('patient_id', 'treatment_date'))

# Check frequencies of all drug types used
table(firstline$drug_code.y)

# List counts of all possible drug combinations per patient (BC only, CC only, and both)
firstline_bc <- subset(firstline, type == "Breast Cancer") %>%
  group_by(patient_id) %>%
  arrange(patient_id, drug_code.y) %>%
  summarize(combination = paste0(drug_code.y, collapse = "-"), .groups = "drop") %>%
  count(combination)

firstline_cc <- subset(firstline, type == "Colon Cancer")%>%
  group_by(patient_id) %>%
  arrange(patient_id, drug_code.y) %>%
  summarize(combination = paste0(drug_code.y, collapse = "-"), .groups = "drop") %>%
  count(combination)

firstline_both <- subset(firstline, type == "both")%>%
  group_by(patient_id) %>%
  arrange(patient_id, drug_code.y) %>%
  summarize(combination = paste0(drug_code.y, collapse = "-"), .groups = "drop") %>%
  count(combination)
```

Solution (Firstline Drug Therapy Combination with highest frequency):

```
   Breast Cancer Only: A and B
    Colon Cancer Only:    C
         Both BC & CC:    D
```

Question 4

The director of the clinic wants to see a nice, presentable table showing the age, sex assigned at birth, geographic region, and stage at diagnosis of the clinic's patients stratified by whether the patient's first diagnosis was Breast cancer or Colon cancer. Build this table.

```r
# Left join Demographics to the diagnosis.date data (first diagnosis)
table <- diagnosis.date %>%
  left_join(demographics, by='patient_id')

# Convert Birthdate to Age
table$age <- floor(age_calc(as.Date(table$birth_date), units = "years"))

# Create Labels for Sex Categories
table$birth_sex <-
  factor(table$birth_sex, levels=c("M","F"),
         labels=c("Male",
                  "Female"))

# Create Labels for Demographic Variables
label(table$age)       <- "Age"
label(table$birth_sex) <- "Sex"
label(table$region   ) <- "Region"
label(table$stage_dx)  <- "Stage at Diagnosis"

# Create Table Caption
caption  <- "Table1. Demographics of Those First Diagnosed with Breast or Colon Cancer"

# Create Table (Alternatively, can use gtsummary for footnotes, p-values, and other clinical table details)
table1(~ age + birth_sex + region + stage_dx | diagnosis, data=table, caption=caption)
```

Solution (table):

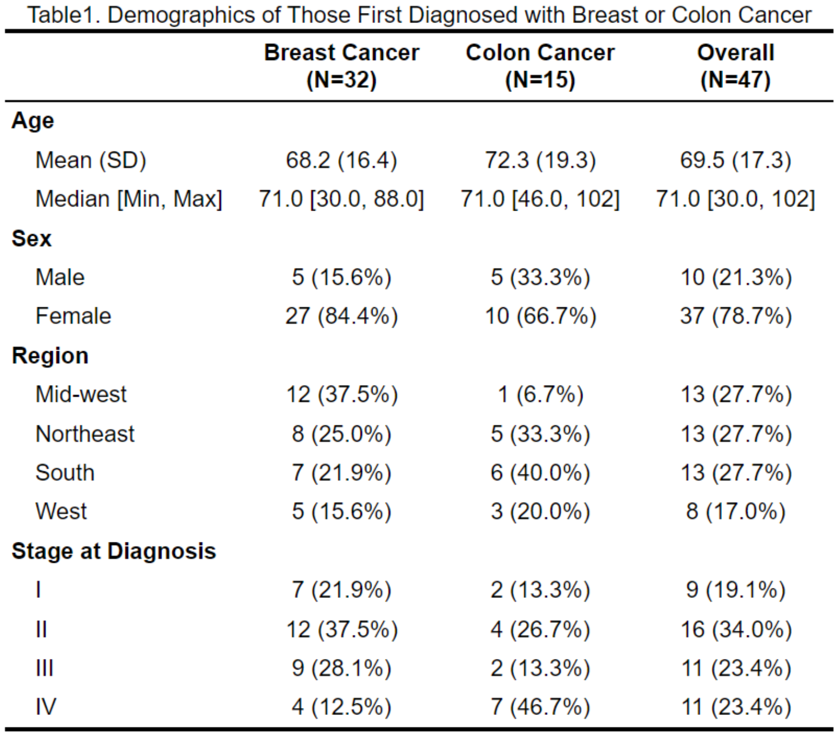

Table1. Demographics of Those First Diagnosed with Breast or Colon Cancer

	Breast Cancer **(N=32)**	**Colon Cancer** **(N=15)**	**Overall** **(N=47)**
Age			
Mean (SD)	68.2 (16.4)	72.3 (19.3)	69.5 (17.3)
Median [Min, Max]	71.0 [30.0, 88.0]	71.0 [46.0, 102]	71.0 [30.0, 102]
Sex			
Male	5 (15.6%)	5 (33.3%)	10 (21.3%)
Female	27 (84.4%)	10 (66.7%)	37 (78.7%)
Region			
Mid-west	12 (37.5%)	1 (6.7%)	13 (27.7%)
Northeast	8 (25.0%)	5 (33.3%)	13 (27.7%)
South	7 (21.9%)	6 (40.0%)	13 (27.7%)
West	5 (15.6%)	3 (20.0%)	8 (17.0%)
Stage at Diagnosis			
I	7 (21.9%)	2 (13.3%)	9 (19.1%)
II	12 (37.5%)	4 (26.7%)	16 (34.0%)
III	9 (28.1%)	2 (13.3%)	11 (23.4%)
IV	4 (12.5%)	7 (46.7%)	11 (23.4%)