BIOST 536: Homework 1

Department of Biostatistics @ University of Washington

Alejandro Hernandez

October 5, 2024

Background

The following questions are related to the evaluation of a clinical trial in patients with acute myelogenous leukemia (AML). The primary endpoint of the clinical trial was induction of complete remission (binary outcome); the two treatments being compared were the newly synthesized anthracycline *idarubicin* and the standard anthracycline agent *daunorubicin*. This assignment ignores certain sequential aspects of the original study and analyzes the data as if investigators always intended to analyze 130 subjects. The complete data and documentation are on the course Canvas site (not publicly accessible).

Descriptive Statistics

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature.

Characteristic	**0**, N = 65	**1**, N = 65
cr	38 (58%)	51 (78%)
sex		
F	30 (46%)	35 (54%)
M	35 (54%)	30 (46%)
age	40 (27, 54)	36 (27, 49)
fab		
0	1 (1.8%)	0 (0%)
1	6 (11%)	13 (21%)
2	15 (26%)	15 (25%)
3	9 (16%)	11 (18%)
4	12 (21%)	8 (13%)
5	13 (23%)	12 (20%)
6	1 (1.8%)	2 (3.3%)
karn		
30	0 (0%)	1 (1.5%)
40	3 (4.6%)	0 (0%)
50	0 (0%)	1 (1.5%)
60	5 (7.7%)	4 (6.2%)
70	6 (9.2%)	11 (17%)
80	27 (42%)	25 (38%)
90	23 (35%)	22 (34%)
100	1 (1.5%)	1 (1.5%)

Characteristic	**0**, N = 65	**1**, N = 65
wbc	17 (3, 70)	12 (3, 42)
plt	62 (36, 126)	50 (32, 78)
hgb	9.45 (8.68, 10.23)	9.20 (8.00, 10.20)
status		
A	16 (25%)	27 (42%)
D	49 (75%)	38 (58%)

Measures of Association

2. Summarize the data in a 2x2 table where outcome D is complete remission and exposure E is treatment group. Estimate the RR, RD, and OR. Which of the three summary measures do you think AML patients would be most interested in?

```
##                     Treatment
## Complete Resmission  0  1
##                   0 27 14
##                   1 38 51
```

The **relative risk** of complete remission given exposure is 1.342. Then, according to our clinical trial, we estimate the likelihood of reaching complete remission to be approximately 1.3 times greater in the arm treated with idarubicin than in the arm treated with daunorubicin. In simple terms: the proportion of patients in our sample who reached complete remission as a part of the new treatment group is approximately 1.3 times greater than the proportion of patients who reached complete remission as part of the standard treatment group.

The **risk difference** of complete remission given exposure is 0.2. We estimate the probability of reaching complete remission to differ between our treatment groups by approximately 20%, with the idarubicin-treated arm having greater chances of completing remission.

The **odds ratio** of complete remission given exposure is 2.588. We estimate the odds of reaching complete remission to be approximately 2.6 times greater in the arm treated with idarubicin than in the arm treated with daunorubicin.

3. Summarize the data in a pair of 2x2 tables as done in Lecture 2, where D is complete remission, E is treatment group, and the covariate is sex.

```
## , , Female = 0
##
##                     Treatment
## Complete Resmission  0  1
##                   0 18  9
##                   1 17 21
##
## , , Female = 1
##
##                     Treatment
## Complete Resmission  0  1
##                   0  9  5
##                   1 21 30
```

Logistic Regression

4. Perform a logistic regression analysis to assess the treatment effect of idarubicin compared to daunorubicin adjusted for sex. In other words, estimate the sex-adjusted OR and present in language suitable for scientific publication.

The **odds ratio** of complete remission given exposure and sex is 2.512. We estimate the odds of reaching complete remission to be approximately 2.5 times greater in the arm treated with idarubicin than in the arm treated with daunorubicin, controlling for sex.

5. Using the subset of data on males, perform a logistic regression analysis to assess the treatment effect of idarubicin compared to daunorubicin for males. Repeat for females.

Among the subset of male patients, the **odds ratio** of complete remission given knowledge of their treatment group is 2.512. We estimate the odds of male patients reaching complete remission to be approximately 2.5 times greater in the arm treated with idarubicin than in the arm treated with daunorubicin, controlling for sex.

Among the subset of female patients, the **odds ratio** of complete remission given knowledge of their treatment group is 2.471. We estimate the odds of female patients reaching complete remission to be approximately 2.5 times greater in the arm treated with idarubicin than in the arm treated with daunorubicin, controlling for sex.

6. You should have found that the sex-adjusted OR you obtained in Q4 is in between the two sex-specific OR you obtained in Q5. Can you explain why this make sense?

7. Fit a logistic regression model with treatment arm, sex, and their interaction. Use the model to estimate the treatment effect in males, and compare to your result to 5(a). Use the model to estimate the treatment effect in females, and compare to your result in 5(b). Comment on the similarity or difference. In general, when you are asked for a point estimate you should include a confidence interval; however, for this problem you are not required to provide confidence intervals.

```
## (Intercept)          tx         fem      tx:fem
##       0.944       2.471       2.471       1.041
```

8.

(a) Write the population attributable risk (as given in Lecture 1) as a function of the rate of exposure P[E] and the relative risk of disease RR.

(b) Suppose smokers have 22 times the risk of dying from lung cancer as non-smokers. Consider a population of 35% smokers. Estimate the PAR for smoking and lung cancer death (point estimate only). Write a sentence presenting and interpreting the PAR.

(c) Suppose smokers have 22 times the risk of dying from lung cancer as non-smokers. Consider a population of 5% smokers. Estimate the PAR for smoking and lung cancer death (point estimate only). Write a sentence presenting and interpreting the PAR.

(d) Comment on the difference between the PAR in (b) and (c).

9. Consider the R script `sim_casecontrolsampling.R` discussed in the first day of class. A statistic not considered is the risk difference RD. Would you expect the RD computed on a case-control sample to estimate the RD in the population? Why or why not? You should be able to answer this question based on the principles already discussed. If you want to, you can modify the my.summary function to include the RD and examine the results. However, this is not required for the homework.

End of report. Code appendix begins on the next page.

Code Appendix

```
# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-')
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# clear environment
rm(list=ls())

# load relevant packages
library(dplyr)      # data frame manipulation
library(knitr)      # table formatting
library(gtsummary)  # "table 1" summary
# library(tidyverse)

# load data
aml <- read.csv("../data/leukemia_data.csv") %>%
  dplyr::mutate_at(vars(tx, sex, eval, cr, status, bmtx, incl), as.factor)
aml$tx <- ifelse(aml$tx=="D", 0, 1)
aml$fem <- ifelse(aml$sex=="M", 0, 1)
aml$cr <- ifelse(aml$cr=="N", 0, 1)

## ===============
## Question 1
## ===============
# create "Table 1" summary
aml %>%
  dplyr::select(tx, cr, sex, age, fab, karn, wbc, plt, hgb, status) %>%
  gtsummary::tbl_summary(by = tx, missing = "no") %>%
  bold_labels()

## ===============
## Question 2
## ===============
# 2x2 table to summarize the number of patients who completed remission
# in each treatment arm
aml %>%
  select('Complete Resmission'=cr, 'Treatment'=tx) %>%
  table()

# prob of complete remission given unexposed (standard treatment)
pRgU <- mean(subset(aml, tx==0)$cr)
# prob of complete remission given exposed (new treatment)
pRgE <- mean(subset(aml, tx==1)$cr)

# relative risk of complete remission given exposure
rr.E <- pRgE / pRgU
# risk difference of complete remission given exposure
rd.E <- pRgE - pRgU
# odds ratio of complete remission given exposure
odds.E <- pRgE / (1 - pRgE)
odds.notE <- pRgU / (1 - pRgU)
```

```r
or.E <- odds.E / odds.notE

## ===============
## Question 3
## ===============
# 2x2 table to summarize the number of patients who completed remission
# in each treatment arm, by sex
aml %>%
  select('Complete Resmission'=cr, 'Treatment'=tx, Female=fem) %>%
  table()

## ===============
## Question 4
## ===============
mod.lr <- glm(cr ~ tx + fem, family="binomial", data=aml)

## ===============
## Question 5
## ===============
mod.lr2 <- glm(cr ~ tx + fem, family="binomial", data=subset(aml, fem==0))
mod.lr3 <- glm(cr ~ tx + fem, family="binomial", data=subset(aml, fem==1))

## ===============
## Question 7
## ===============
mod.lr3 <- glm(cr ~ tx*fem, family="binomial", data=aml)
round(exp(coef(mod.lr3)), 3)

## ===============
## Question 8
## ===============


## ===============
## Question 9
## ===============
```

End of document.