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	**Daunorubicin**, N = 65	**Idarubicin**, N = 65
cr		
Did not complete remission	27 (42%)	14 (22%)
Completed remission	38 (58%)	51 (78%)
sex		
Female	30 (46%)	35 (54%)
Male	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%)
Unknown	8	4
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%)

Characteristic	**Daunorubicin**, N = 65	**Idarubicin**, N = 65
80	27 (42%)	25 (38%)
90	23 (35%)	22 (34%)
100	1 (1.5%)	1 (1.5%)
wbc	17 (3, 70)	12 (3, 42)
Unknown	1	0
plt	62 (36, 126)	50 (32, 78)
Unknown	1	0
hgb	9.45 (8.68, 10.23)	9.20 (8.00, 10.20)
Unknown	1	0
eval		
N	5 (7.7%)	5 (7.7%)
Y	60 (92%)	60 (92%)
status		
A	16 (25%)	27 (42%)
D	49 (75%)	38 (58%)
bmtx		
N	59 (91%)	57 (88%)
Y	6 (9.2%)	8 (12%)
incl		
N	20 (31%)	20 (31%)
Y	45 (69%)	45 (69%)

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?

Table 2: Contingency table of patients who completed remission in each treatment arm

	Daunorubicin	Idarubicin
Did not complete remission	27	14
Completed remission	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.

I expect patients with acute myelogenous leukemia to be most interested in the risk difference associated with treatment, as it provides the most direct and meaningful interpretation in terms of the absolute change in risk due to treatment.

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.

Table 3: Contingency table of female patients who completed remission in each treatment arm

	Daunorubicin	Idarubicin
Did not complete remission	9	5
Completed remission	21	30

Table 4: Contingency table of male patients who completed remission in each treatment arm

	Daunorubicin	Idarubicin
Did not complete remission	18	9
Completed remission	17	21

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.

We estimate that the odds of complete remission from acute myelogenous leukemia (AML) is 2.51 times greater for a population treated with idarubicin compared to a population of the same sex treated with daunorubicin (95% CI for odds ratio: 1.16 - 5.63).

This analysis of our sample suggests the newly synthesized anthracycline, idarubicin, has a statistically significant, positive effect on the induction of complete AML remission. However, our confidence interval indicates that observing a weak effect within a patient population would not be unusual.

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.

We estimate that the odds of complete AML remission is 2.47 times greater for a male population treated with idarubicin compared to a male population treated with daunorubicin (95% CI for odds ratio: 0.9 - 7.1).

Comparing two treated female populations, we estimate the odds of complete AML remission to be 2.57 times greater for females receiving idarubicin than those receiving daunorubicin (95% CI for odds ratio: 0.77 - 9.41).

Neither analyses of these subsamples suggest the idarubicin treatment has a statistically significant effect on the induction of complete remission of AML. Still, our confidence interval indicates that observing a strong effect within either group of patients would not be unusual.

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?

The subgroups that fashioned the two conditional logistic regression models form a complete partition of the sample that fit the marginal logistic regression model; because logistic regression models the proportion of a population, the estimated proportion of the overall population will be an aggregate of the estimated proportions of the subpopulations.

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.

We estimate that the odds of complete AML remission is 2.47 times greater for a male population treated with idarubicin compared to a male population treated with daunorubicin (95% CI for odds ratio: 0.15 - 44.46).

Comparing two treated female populations, we estimate the odds of complete AML remission to be 2.57 times greater for females receiving idarubicin than those receiving daunorubicin (95% CI for odds ratio: 0.77 - 9.41).

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.

$$PAR = \frac{p_e\{\mathrm{P}(D \mid E) - \mathrm{P}(D \mid \overline{E})\}}{p_e\mathrm{P}(D \mid E) + (1 - p_e)\mathrm{P}(D \mid \overline{E})}$$

Expanding the denominator, then dividing the numerator and denominator by $\mathrm{P}(D \mid \overline{E})$, produces

$$PAR = \frac{p_e\{\mathrm{P}(D \mid E) - \mathrm{P}(D \mid \overline{E})\}}{p_e\{\mathrm{P}(D \mid E) - \mathrm{P}(D \mid \overline{E})\} + \mathrm{P}(D \mid \overline{E})}$$

$$PAR = \frac{p_e\{RR - 1\}}{p_e\{RR - 1\} + 1}$$

Where RR denotes the relative risk of disease associated with the exposure. Note that if we have conducted a case-control study of a rare disease D, the RR associated with exposure is approximately equal to the odds ratio (OR) associated with exposure. Then,

$$PAR \approx \frac{p_e\{OR - 1\}}{p_e\{OR - 1\} + 1}$$

(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.

$PAR = \frac{(0.35)\{22-1\}}{(0.35)\{22-1\}+1} = 0.880$

In a population of 35% smokers, the proportion of the overall risk of fatal lung cancer that is due to smoking is 88%.

(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.

$PAR = \frac{(0.05)\{22-1\}}{(0.05)\{22-1\}+1} = 0.512$

In a population of 5% smokers, the proportion of the overall risk of fatal lung cancer that is due to smoking is 51.2%.

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

If smokers have a higher risk of dying from lung cancer than non-smokers, then in populations with more smokers, a larger share of lung cancer deaths will be due to smoking. We found this to be true when comparing two the populations of (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.

We would not expect the a population's risk difference associated with exposure to be estimated by a case-control sample because such studies sample a fixed number of cases and controls, which does not reflect the actual prevalence of disease.

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(sjPlot)     # model coefficient table
# library(tidyverse)

# load data
aml <- read.csv("../data/leukemia_data.csv") %>%
  # read in select variables as factors
  dplyr::mutate_at(vars(tx, sex, eval, cr, status, bmtx, incl), as.factor)
# create boolean variables
aml$tx2 <- ifelse(aml$tx=="D", 0, 1)
aml$sex2 <- ifelse(aml$sex=="M", 0, 1)
aml$cr2 <- ifelse(aml$cr=="N", 0, 1)
# rename levels of select factors
levels(aml$tx) <- c("Daunorubicin", "Idarubicin")
levels(aml$sex) <- c("Female", "Male")
levels(aml$cr) <- c("Did not complete remission", "Completed remission")

## ===============
## Question 1
## ===============
# create "Table 1" summary
aml %>%
  dplyr::select(tx, cr, sex:eval, status, bmtx, incl) %>%
  gtsummary::tbl_summary(by = tx) %>%
  bold_labels()

## ===============
## Question 2
## ===============
# 2x2 table to summarize the number of patients who completed remission
# in each treatment arm
table(aml$cr, aml$tx) %>%
  knitr::kable(caption = "Contingency table of patients who completed remission
               in each treatment arm")

# prob of complete remission given unexposed (standard treatment)
pRgU <- mean(subset(aml, tx2==0)$cr2)
# prob of complete remission given exposed (new treatment)
pRgE <- mean(subset(aml, tx2==1)$cr2)
```

```r
# 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)
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
subset(aml, sex == "Female", select = c(cr, tx)) %>%
  table() %>%
  kable(caption = "Contingency table of female patients who completed remission
        in each treatment arm")

subset(aml, sex == "Male", select = c(cr, tx)) %>%
  table() %>%
  kable(caption = "Contingency table of male patients who completed remission
        in each treatment arm")

## ===============
## Question 4
## ===============
mod.lr <- glm(cr ~ tx + sex, family = binomial(), data=aml)
exp.coef <- exp(coef(mod.lr))
exp.confint <- exp(confint(mod.lr))

# print model coefficients with sjPlot (not displayed in PDF)
sjPlot::tab_model(mod.lr, title = "Logistic regression of complete remission",
                  show.r2 = F, show.aic = T)

## ===============
## Question 5
## ===============
mod.lr2 <- glm(cr ~ tx, family = binomial(), data=subset(aml, sex2==0))
exp.coef2 <- exp(coef(mod.lr2))
exp.confint2 <- exp(confint(mod.lr2))

mod.lr3 <- glm(cr ~ tx, family = binomial(), data=subset(aml, sex2==1))
exp.coef3 <- exp(coef(mod.lr3))
exp.confint3 <- exp(confint(mod.lr3))

# print model coefficients with sjPlot (not displayed in PDF)
tab_model(mod.lr2, title = "Logistic regression of complete remission among
          male patients", show.r2 = F, show.aic = T)
tab_model(mod.lr3, title = "Logistic regression of complete remission among
          female patients", show.r2 = F, show.aic = T)

## ===============
```

```r
## Question 7
## ===============
mod.lr4 <- glm(cr ~ tx*sex, family=binomial(), data=aml)
coef4 <- coef(mod.lr4)
confint4 <- confint(mod.lr4)

# print model coefficients with sjPlot (not displayed in PDF)
tab_model(mod.lr4, title = "Logistic regression of complete remission",
          show.r2 = F, show.aic = T)
```

End of document.