**Biostat/Epi 536 2024**

**HW 1 (9 questions)**

The following questions are related to the evaluation of a clinical trial in patients with acute myelogenous leukemia (AML). The data and documentation are on the course CANVAS site. The primary endpoint of the clinical trial was induction of complete remission (binary outcome).

You will see in the data description that there are some complicating factors in how the clinical trial was executed. For this homework, we will ignore the sequential aspect of the clinical trial and analyze the data as if the investigators had always planned on getting data on 130 subjects. We will use an intent-to-treat approach, which means that our predictor of interest is the groups to which patients were randomized (regardless whether the treatment received matches the treatment assigned).

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature.
2. Summarize the data in a 2x2 table where D is complete remission and E is treatment group. Estimate the RR, RD, and OR (you do not have to produce confidence intervals). Which of the three summary measures do you think AML patients would be most interested in: the OR, RR, or RD?
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.
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.
5. (a) Using the subset of data on males, perform a logistic regression analysis to assess the treatment effect of idarubicin compared to daunorubicin for males.

(b) Repeat for females.

1. 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?
2. 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.

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.