Data Analysis - 2016 First Year Qualifying Exam
Department of Statistics, University of California, Irvine

Handed out: **Monday, June 20, 2016 at 12:00pm** Due: **Friday, June 24, 2016 at 5:00pm**

Turning In Your Exam: Email your complete solution to BOTH Rosemary Busta (rbusta@ics.uci.edu) and Dan Gillen (dgillen@uci.edu) by 5pm on Friday, June 24. LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.

1 Background

In the United States, cancer is the leading cause of death in people under age 75 years, and colorectal cancer is the second most common cause of cancer deaths. Colorectal cancer may be prevented by removal of precursor adenomas found during screening sigmoidoscopy or colonoscopy. Adenomas are benign tumors found in the colon, are are a type of colon polyp. Although adenomas are benign, over time they may transform to become malignant, at which point they are called adenocarcinomas. While removal of adenomas are a way to prevent colorectal cancer, screening rates in the general population of individuals most at risk are variable and range from 30% to 90% depending highly on reimbursement policies. As such, many adenomas may go undetected via screening and hence other preventive measures are in need.

Diet and inflammation are two factors that have been associated with risk of colorectal cancer, and a series of clinical trials have been conducted to test the efficacy of individual dietary supplements or anti-inflammatory agents to prevent the incidence or recurrence of colon polyps. Unfortunately, these trials have not translated into significant changes in medical practice for prevention or management of colon cancer for a variety of reasons, including lack of efficacy, unacceptable toxicities, and the availability of competing strategies for risk reduction.

Studies in rodent models have shown that combination chemoprevention strategies are often more effective than those using individual agents. One particular enzyme-activated, inhibitor of ornithine decarboxylase that we will refer to as for the purposes of this analysis as D-CARB, has been identified as a potent inhibitor of intestinal and colon carcinogenesis in animal models, especially in combination with nonsteroidal anti-inflammatory drugs such as aspirin or ibuprofen. D-CARB and aspirin have also been shown to interact to prevent the growth and viability of human colon cancer cells in in vitro models. However, D-CARB has also been associated with potential adverse effects in mouse models. Namely, it has been observed and reported that treatment with D-CARB may negatively impact hearing and this impact may be dependent on the cumulative duration of exposure to D-CARB. Studies that have reported hearing loss associated with D-CARB use have found the absolute change in hearing to be consistent across multiple sound frequencies.

Given the early mouse and lab work in D-CARB, a multi-site double-blind randomized clinical trial was conducted to quantify the effect of D-CARB treatment on reducing the rate of development of adenomas over 36 months. $N = 364$ high risk patients with a history of resected (¿ or =3 mm) adenomas were consented for participation and randomly assigned in a 1:1 fashion to receive oral D-CARB 600 mg once daily or a matched placebo for 36 months. The randomization was stratified on the basis of the use (defined as 81 mg daily or 325 mg twice weekly) or nonuse of low-dose aspirin at study entry. This stratification ensure that an equal proportion of aspirin users are present in each treatment arm. Patients received annual sigmoidoscopies to determine the development of adenomas. The primary outcome for trial was the total number of adenomas detected in a patient over the 36 month observation period.

To assess the impact of D-CARB on hearing loss, a randomly chosen subset of patients ($N = 184$) were asked to take audiology tests at baseline (the time of randomization), approximately 18 months after the start of treatment, and approximately 36 months after the start of treatment. The timing of these visits could vary widely across patients. At each visit, sound frequencies tested were 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz and 8000 Hz (note that normal speech range is 500-3000 Hz). For each frequency, the test proceeds by stepping up the sound

by 5 decibels (dB) until the patient indicates that the sound can be hear. Thus a lower dB scores considered better hearing performance. At each visit, the test was repeated in both the right and left ear of the patient. For some patients, not all frequencies had dB values recorded for one or both ears due to a variety of reasons.

2 Scientific Goals

Using the data obtained from the clinical trial, your job is to address the following aims:

1. Determine if D-CARB is effective in reducing polyp development over 36 months.
2. Determine if D-CARB is more or less effective in patients that use aspirin daily.
3. Quantify the potential impact of treatment with D-CARB on hearing loss.

3 General Instructions

You are to analyze the data (see below) to best address the scientific goals stated above. You should use appropriate statistical methods for estimating and quantifying uncertainty in associations. Your final analysis should be presented in the form of a brief report (no more than 10 pages including relevant tables and figures). You may place additional information (eg. relevant diagnostic plots) in an Appendix if you feel it necessary. The report should (at minimum) consist of the following sections:

1. Abstract - A brief summary of your basic findings.
2. Introduction - Background on the scientific problem, an introduction to the problem at hand, and what is to be addressed.
3. Statistical Methods - A clear discussion and justification of the methods you have used to analyze the data and the modeling strategy that you employed.
4. Results - A presentation of the results of your analysis that includes relevant and properly formatted tables and figures as well as complete and precise interpretations of your analytic findings.
5. Discussion - A synopsis of your findings, what they have achieved with respect to the scientific goals, any limitations your analysis may suffer from, and possible future directions to better achieve the scientific goals you set out to accomplish.

Your report should be well-written, succinct, and to the point! It should be written in a language that is understandable to the broad scientific community while precisely interpreting your finding. Thje discussion of statistical methods should be more technical than that provided to a non-statistical audience given the purpose of the report. It should be complete but brief - free of garbage and not-so-relevant material. You are encouraged to use relevant and well-formatted tables, plots and figures to help explain your findings. You may use any written references for this problem that you wish, **but you cannot communicate (talk, email, etc) with anyone about your analysis**.

4 Available Data

There are two datasets available for use in this analysis (`demo_outcome_data` and `audio_long`). The first dataset, `demo_outcome_data`, contains all baseline demographic, treatment, and primary response (number of adenomas

observed over 3 years) data for all ($N = 364$) patients enrolled in the trial. The second dataset, `audio_long`, contains the longitudinal audiology data for each subject enrolled in the hearing substudy ($N = 184$). These data formatted in a "long" format in which the hearing results for each patient visit along with the date of the test are placed on a separate row. Note that the the patients in the audiology substudy are a subset of those in the main trial and hence their demographic and treatment data can be merged by using the unique study ID for each patient (`study_no`). The data are available as an ASCII representation of an R file and can be loaded into R with the following commands:

```
demo_outcome_data <- dget( "http://www.ics.uci.edu/~dgillen/FYQualExam2016/demo_outcome_data" )
audio_long <- dget( "http://www.ics.uci.edu/~dgillen/FYQualExam2016/audio_long" )
```

A brief description of the variables in each datsaset is given below:

Table 1: Variable description for the `demo_outcome_data` dataset.

Column	Variable Name	Description
1	`study_no` :	Unique patient ID
2	`site` :	Study site where patient was recruited
3	`tx` :	Assigned treatment (1=D-CARB, 0=Placebo)
4	`asa_usage` :	Daily aspirin usage (1=yes, 0=no)
5	`age` :	Patient age in years at time of randomization
6	`sex` :	Patient gender (factor variable with levels Female and Male)
7	`ethnic` :	Patient race/ethnicity (factor variable indicating race/ethnicity identification)
8	`adenomas` :	Number of adenomas observed over 3 years

Table 2: Variable description for the `audio_long` dataset.

Column	Variable Name	Description
1	`study_no` :	Unique patient ID
2	`days_on_tx` :	Total days patient was on treatment over course of study
3	`meas_date` :	Date of the audiology test (in R `Date` format)
4	`left_250` :	Hearing detection (dB) at 250 Hz (left ear)
5	`left_500` :	Hearing detection (dB) at 500 Hz (left ear)
6	`left_1000` :	Hearing detection (dB) at 1000 Hz (left ear)
7	`left_2000` :	Hearing detection (dB) at 2000 Hz (left ear)
8	`left_3000` :	Hearing detection (dB) at 3000 Hz (left ear)
9	`left_4000` :	Hearing detection (dB) at 4000 Hz (left ear)
10	`left_6000` :	Hearing detection (dB) at 6000 Hz (left ear)
11	`left_8000` :	Hearing detection (dB) at 8000 Hz (left ear)
12	`right_250` :	Hearing detection (dB) at 250 Hz (right ear)
13	`right_500` :	Hearing detection (dB) at 500 Hz (right ear)
14	`right_1000` :	Hearing detection (dB) at 1000 Hz (right ear)
15	`right_2000` :	Hearing detection (dB) at 2000 Hz (right ear)
16	`right_3000` :	Hearing detection (dB) at 3000 Hz (right ear)
17	`right_4000` :	Hearing detection (dB) at 4000 Hz (right ear)
18	`right_6000` :	Hearing detection (dB) at 6000 Hz (right ear)
19	`right_8000` :	Hearing detection (dB) at 8000 Hz (right ear)