Project Proposal

[redacted]

[redacted] 2022

1 Group members

[redacted]

2 Problem Statement

We want to understand how much smoking increases the risk of certain cardio-vascular related ailments:

1. Coronary heart disease (CHD)
2. Hypertension
3. Stroke
4. Myocardial Infarction (MI)

These are all serious medical ailments, so identifying their causes is of interest in order to understand how to prevent them. Medical professionals and everyday people trying to live a healthy lifestyle are parties interested in this question.

3 Causal Questions or Hypotheses

(a) **Causal Question**: Let A_1 be whether or not an individual is a smoker, let A_2 be the number of cigarettes he or she smokes per day, and let Y_1, Y_2, Y_3, Y_4 be the development of CHD, hypertension, stroke, and MI respectively. Let $\Theta = \{A_1, A_2\}$ and let $\Psi = \{Y_1, Y_2, Y_3, Y_4\}$. We are interested in the all combinations of these treatments and outcomes represented by $\Theta \times \Psi$.

(b) **Causal Estimate**: For every treatment-outcome pair in $A_1 \times \Psi$, we can look at the causal risk ratio $E[Y_i^{a=1}]/E[Y_i^{a=0}]$, or the expected rate of ailment Y_i if one smokes divided by the expected rate of ailment Y_i if one doesn't smoke. For every treatment-outcome pair in $A_2 \times \Psi$, we can look at simply the $E[Y_i^a]$, which is a function of a even after we fix i. We will attempt to approximate this function which represents the expected rate of ailment Y_i as a function of values of treatment A_2.

(c) **Comments**: Any of these causal questions would be interesting. The subset $A_1 \times \Psi$ is less complex as we are dealing with a dichotomous treatment; however, it may also be less interesting. We postulate that the distinction between which cardiovascular ailment we measure as an outcome is not as important as they are likely constructed from similar graphical models. Therefore, it may be plausible to simply exchange CHD for MI and so forth. This is not a guarantee but a prediction.

4 Dataset(s)

We will use the Framingham Heart Study Longitudinal Dataset, which was provided by the instructor.

(a) **Background and Contents**: This study took began in 1948 with 5,209 initial subjects. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes, specifically with occurrences of Coronary Heart Disease (CHD), Myocardial Infarction (MI), Stroke, and Hypertension. Our data set is a subset of this data on 4,434 participants, followed for 18 years (or death/lack of followup, whichever is first) and followed up with every 6 years.

(b) **Limitations**: The data has several limitations:

1. Missing data. Some participants died or did not follow-up. Thus, the data is not complete.
2. The variable HYPERTEN is biased upwards since defining Hypertensive requires exam participation and bias can therefore occur. Subjects attending exams regularly have a greater opportunity to be defined as hypertensive. Subjects not attending exams would be assumed to be free of hypertension. Since Hypertension is highly prevalent, this misclassification could potentially be large.
3. Limited behavioral data included. Most of the variables appear to be outcomes (types of cardiovascular outcomes) or mediators (diabetes).

(c) **Format**: This data is longitudinal and has a new entry for every visit (3 visits: one every six years), and thus one individual may have up to three entries. Additionally, the variables that track whether or not a cardiovascular outcome occurs and when (ex. mi_fchd and timeifc) are the same across all instances of a persons visit regardless of when it occurred. Only the first incident is recorded. Additionally, the follow-ups in the data are every 6 years. Along with the occurrences of cardiovascular outcomes, the data records demographic information such as sex and age as well as health/behavioral information, such as whether or not the participant is a smoker, his/her BMI, etc... Most data is dichotomous or at least categorical. Missing data is notated with a "." The data could be manipulated so that every row corresponded to one individual and the columns were all the features, but this would entail tripling the number of features (one set for each possible visit).

(d) **DataFrame Preview**: Here are 20 example entries with 8 selected variables: RANDID, PERIOD, TIME, CURSMOKE, CIGPDAY, PREVCHD, ANYCHD, TIMECHD.

	RANDID	PERIOD	TIME	CURSMOKE	CIGPDAY	PREVCHD	ANYCHD	TIMECHD
0	2448	1	0	0	0.0	0	1	6438
1	2448	3	4628	0	0.0	0	1	6438
2	6238	1	0	0	0.0	0	0	8766
3	6238	2	2156	0	0.0	0	0	8766
4	6238	3	4344	0	0.0	0	0	8766
5	9428	1	0	1	20.0	0	0	8766
6	9428	2	2199	1	30.0	0	0	8766
7	10552	1	0	1	30.0	0	0	2956
8	10552	2	1977	1	20.0	0	0	2956

9	11252	1	0	1	23.0	0	0	8766
10	11252	2	2072	1	30.0	0	0	8766
11	11252	3	4285	1	30.0	0	0	8766
12	11263	1	0	0	0.0	0	1	5719
13	11263	2	2178	0	0.0	0	1	5719
14	11263	3	4351	0	0.0	0	1	5719
15	12629	1	0	0	0.0	0	1	373
16	12629	2	2212	0	0.0	1	1	373
17	12806	1	0	1	20.0	0	0	8766
18	12806	2	2170	1	30.0	0	0	8766
19	12806	3	4289	1	30.0	0	0	8766

...

(e) **Variable Overview**: The example entries include 8 selected variables.

i. RANDID is the unique identification number for each participant, which is a discrete integer ranging from 2448 to 9999312.

ii. PERIOD is the current examination cycle, which is discrete and is either 1, 2 or 3.

iii. TIME is the number of days since the baseline exam, which is a discrete integer ranging from 0 to 4854.

iv. CURSMOKE is a binary variable, where 1 represents the participant is smoking during the current examination cycle and 0 for not smoking.

v. CIGPDAY is the number of cigarettes the participant smoked each day, which is a discrete integer ranging from 0 to 90.

vi. PREVCHD is a binary variable, representing whether the participant has the prevalent Coronary Heart Disease during the current examination cycle.

vii. ANYCHD is a binary variable, representing whether the participant has the incident Coronary Heart Disease during the current examination cycle.

viii. TIMECHD is the time of the first ANYCHD event during the followup measured as the number of days since the baseline exam, which is a discrete integer and has a maximum value of 8766 among the entries.

(f) **Missing Data**: There are many potential unmeasured confounders such as the participant's socioeconomic status and profession, whether people around the participant smoke or not, etc., which will have impacts on both the participant's likelihood to smoke (CURSMOKE and CIGPDAY) and health outcomes (e.g., ANYCHD). The lack of measurements for confounders could be particularly problematic because there are no obvious mediators between smoking and health outcomes in the dataset that can be used to determine the causal relation.

5 Expectations and Concerns

We may want to apply the structural learning techniques with the graphical model from the class, practice the use of such a computational tool on the real-world dataset, and explore how different approaches perform on the dataset. With the computed graphical model, we can verify our assumptions of the relations between variables and investigate the causal effect between the treatments and the outcomes.

The main concern about this particular data is the lack of obvious mediators between our selection of treatments and outcomes. As a result, the graphical model may not be able to capture the biases from many unobserved confounders for the causal relation we intend to study.