Project Update

CS396-4 Causal Inference

[redacted] 2022

1 Group members

[redacted]

2 Causal graph

The team is using the Framingham Heart Study to study the effect of cigarette smoking on heart disease. For some background, participant clinic data was collected during three examination periods (check-ups), approximately 6 years apart, from roughly 1956 to 1968. After these check-ups, there was a follow-up period and in total, each participant was followed for a total of 24 years for the outcome of the following events: Angina Pectoris, Myocardial Infarction, Atherothrombotic, Infarction or Cerebral Hemorrhage (Stroke) or death. The response variable whose effect the team is measuring is whether or not heart disease is detected in a participant. Since, this is a longitudinal study which has three separate check-ups and then a follow-up period, there are four distinct categories of relevant participants as they relate to this causal outcome. These categories are determined by when a participant first develops heart disease. They are as follows:

A. Participants who have heart disease at the first check-up

B. Participants whose first record of heart disease occurs at the second check-up

C. Participants whose first record of heart disease occurs at the third check-up

D. Participants whose first record of heart disease occurs at the follow-up after all three check-ups.

E. Participants who have not yet had a record of heart disease by the conclusion of all three check-ups and the follow-up.

Participants in group A are not included in the analysis because they already had heart disease at the start of the study. Participants in groups B, C, D and E are included in the study, but the causal effect of smoking on heart disease is modeled differently for each group. Once a person develops heart disease, they are excluded from the model which models heart disease at the next check-up or the follow-up. The graphical models are included and labeled below. The essential idea behind these graphical models is that there is a treatment, an outcome, some mediators, and some confounders for each check-up. If a participant has not recorded heart disease, at a check-up then the treatment, mediators, and counfounders of this model all effect the set of treatments, mediators, and confounders for the next check-up. Here is an analysis of the remaining variables and their meanings:

A#: The treatment for check-up (#+1):

1. CURSMOKE (Current Cigarette Smoking at Exam)

M#: The mediators for check-up (#+1):

1. SYSBP (Systolic Blood Pressure (mmHg))
2. DIABP (Diastolic Blood Pressure (mmHg))
3. BPMEDS (Use of Anti-hypertensive Medication at Exam)
4. TOTCHOL (Serum Total Cholesterol (mg/dL))
5. HDLC (HDL Cholesterol (mg/dL))
6. LDLC (LDL Cholesterol (mg/dL))
7. GLUCOSE (Casual Serum Glucose (mg/dL))

C#: The confounders for check-up (#+1):

1. SEX
2. AGE
3. EDU

Y#: The outcome for check-up (#+1):

1. PREVCHD (Coronary Heart Disease Recorded before Exam) – For Y0, Y1, and Y2
2. ANYCHD (Coronary Heart Disease Recorded in Follow-Up) – For Y2

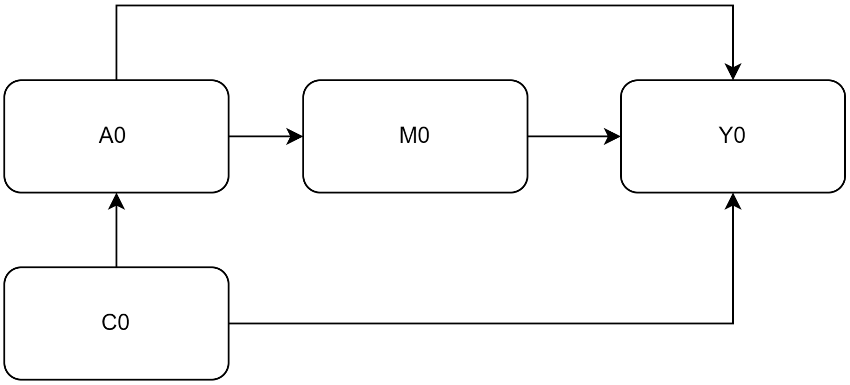

Figure 1: Graphical Model for Comparing Group B with Groups C, D and E

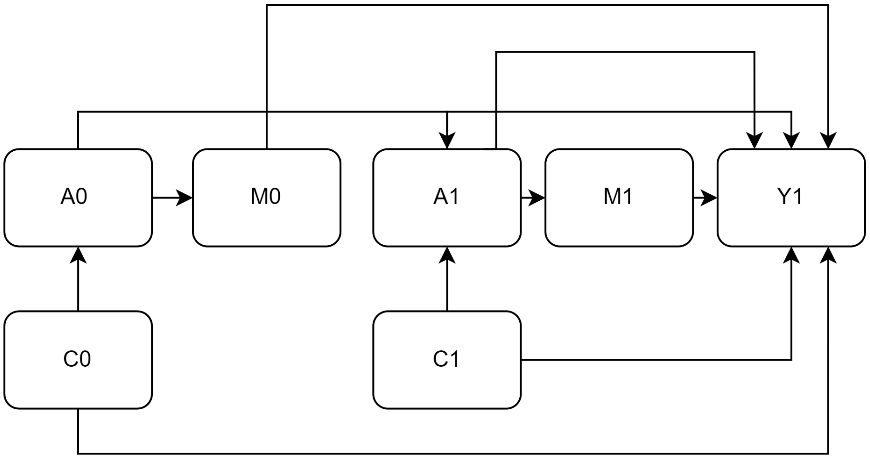

Figure 2: Graphical Model for Comparing Group C with Groups D and E

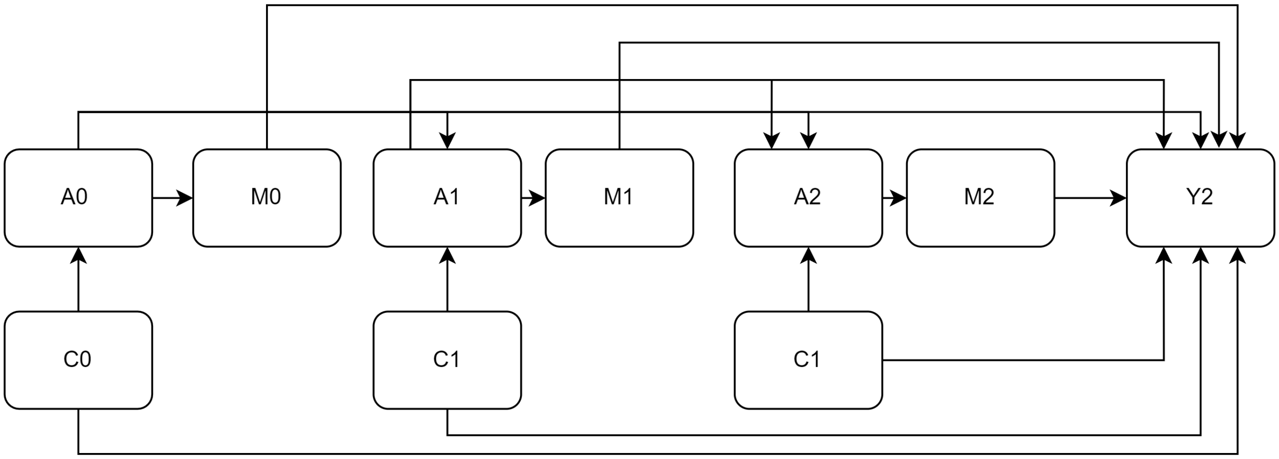

Figure 3: Graphical Model for Comparing Group D with Group E

3 Counterfactual function

The counterfactual functions for different periods are respectively as follows:

$$E(Y0|A0 = a0) = \sum_{C0} E(Y0|C0, A0 = a0)P(C0) \tag{1}$$

$$E(Y1|A0 = a0, A1 = a1) = \sum_{C0,C1} E(Y1|C0, C1, A0 = a0, A1 = a1)P(C1)P(C0) \tag{2}$$

$$\begin{aligned} E(Y2|A0 = a0, A1 = a1, A2 = a2) = \sum_{C0,C1,C2} & E(Y2|C0, C1, C2, A0 = a0, A1 = a1, A2 = a2) \\ & *P(C2)P(C1)P(C0) \end{aligned} \tag{3}$$

3.1 Assumptions

Notice that the above counterfactual functions are all backdoor estimators and thus the common assumptions that come with using a backdoor estimator hold. Thus, we the team is assuming consistency, conditional exchangeability, and the conditional independence of the treatments (A0,A1, and A2) with the outcomes (Y0,Y1, and Y2) given the confounders (C0, C1, and C2).

4 Estimation and Interpretation

The team is interested in finding the causal risk ratio between participants who reported smoking at all check-ins (up to the present) to those who reported not smoking at all check-ins (up to the present). Specifically, with A# = 1 signifying the participant reporting that they smoke at check-in (#+1), and A# = 0 signifying the participant reporting that they don't smoke at check-in (#+1), the team found the following ratios:

1. $E(Y0|A0 = 1)/E(Y0|A0 = 0) = 0.03753/0.02899 = 1.30$ (n=3053)
2. $E(Y1|A0 = 1, A1 = 1)/E(Y1|A0 = 0, A1 = 0) = 0.06061/0.05581 = 1.09$ (n=2952)
3. $E(Y2|A0 = 1, A1 = 1, A2 = 1)/E(Y2|A0 = 0, A1 = 0, A2 = 0) = 0.18144/0.15987 = 1.13$ (n=2782)

These ratios give us the percent increase in Coronary Heart Disease caused by smoking. To calculate these ratios, the team used a backdoor estimator as described in the equations in Part 3. The team used linear regression to determine the internal expectations within the summands ($E(Y2|C0, C1, C2, A0 = a0, A1 = a1, A2 = a2)$, $E(Y1|C0, C1, A0 = a0, A1 = a1)$, and $E(Y0|C0, A0 = a0)$). The team then calculated the remaining probabilities from the data and computed the sums. As seen above,

4.1 Uncertainty

The corresponding 95% confidence intervals for different periods are as follows:

1. $E(Y0|A0 = 1)/E(Y0|A0 = 0) : [0.86990, 1.92701]$ (n=3053)
2. $E(Y1|A0 = 1, A1 = 1)/E(Y1|A0 = 0, A1 = 0) : [0.78926, 1.46820]$ (n=2952)
3. $E(Y2|A0 = 1, A1 = 1, A2 = 1)/E(Y2|A0 = 0, A1 = 0, A2 = 0) : [0.95971, 1.36002]$ (n=2782)

4.2 Interpretation

The team sees that smoking increases the risk of smoking across the board (regardless of how years smoked). It appears that the causal effect is highest in the short-term with a near 30% increase and over time decreasing to around 10%. Thus, individuals should choose to not smoke if they desire to decrease their risk of heart disease and government policies should discourage smoking. However, none of the results are significant at a 95% confidence level.

5 Next steps for the project

In the next few weeks, there are some challenges the team would like to overcome.

5.1 One additional challenge

Since the Framingham Dataset is longitudinal it is natural that patient's are lost over time for a multitude of reasons, including: death, lack of interest, and moving away. Some patients thus do not report to all check-ins. In our analysis thus far, only patients with complete data were considered. The team will first attempt to causally derive the missing information under a Missing At Random (MAR) model. If this is not possible, the team will consider another form of missing value amputation.

5.2 Ask for feedback

An area of difficulty for the team was deciding on how to integrate the multiple effects for patients. Since there were three check-ins, there are three outcome variables (Y0, Y1, and Y2). The team wonders how these might be combined into a simpler outcome. Additionally, the team is unsure how to get a 95% confidence interval that is significant.