252E Final Project Part 2: Relationship between Frequency of Care and Progression to Hepatocellular Carcinoma among Chronic Hepatitis C Patients

Stephanie Holm and Shelley Facente

11/14/2019

1. Background

a. Brief Introduction

b. Description of our Dataset

We have access to a dataset of chronic Hepatitis C patients currently receiving care in the UCSF system. There are 2297 patients, who were adults by the start of 2011 and have been seen in a variety of primary care clinics and the hepatology (liver) clinic between 2011 and 2018. These data come from a query of Apex, the UCSF-specific build of the electronic medical record system Epic.

2. Roadmap

a. Causal Model

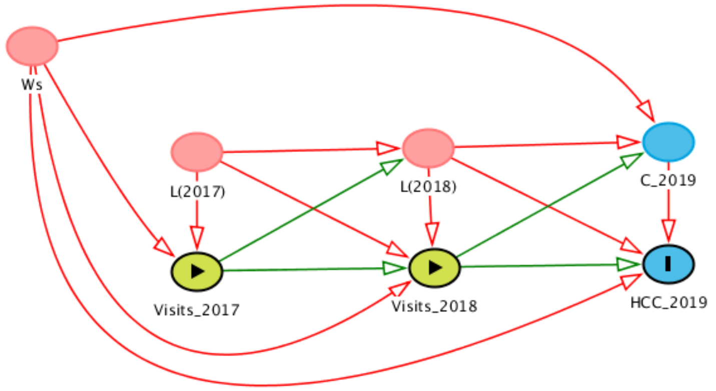
i. Defining Our Variables

Exposures: We have data on the annual number of clinical visits that each chronic HCV patient had in the UCSF system annually from 2011 through 2018.

Outcome: The outcome is diagnosis of hepatocellular carcinoma (HCC), which occured in 361 patients.

Covariates: Annual FIB-4 score (a validated measure used for prediction of cirrhosis, which uses age, platelet count and liver transaminases), biologic sex, race, insurance type (mediCal vs private- a surrogate of socioeconomic status.)

ii. Our DAG



We intend to do this analysis using 9 years worth of data, but in order to simplify the DAG, we present only the final 3 years (two of exposure data and a final year of outcome) here. This DAG includes both a C value indicating whether or not the patient had the outcome measured by the end of the study.

iii. Our Structural Causal Model, $O = (W, C(t), \Delta(t), L(t), Y(t), A(t))$

This is survival data with missingness and censoring, where:

- W is the baseline covariates (race, sex and SES)
- L(t) is the set of covariates (most recent FIB-4 score, years since last FIB-4) at time t
- A(t) is the exposure (number of visits)
- C(t) is an indicator of being censored at the final timepoint (1 means they were censored)
- Y(t) is the outcome (an indicator of HCC diagnosis)

$U = (U_{L(t)}, U_{A(t)}, U_{C(t)}, U_{Y(t)},), t = 1, 2, 3, 4, 5 \sim P_U$

Structural Equations, F:

$$\begin{aligned}
W &= f_W(U_W) \\
L(1) &= f_{L(1)}(U_{L(1)}) \\
A(1) &= f_{A(1)}(W, L(1), U_{A(1)}) \\
L(t) &= f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}) \\
A(t) &= f_{A(t)}(W, \bar{L}(t), \bar{A}(t-1), U_{A(1)}) \\
C(t) &= f_{C(t)}(W, \bar{A}(t-1), \bar{Y}(t-1)) \\
Y(t) &= f_{Y(t)}(W, C(t), \bar{L}(t-1), \bar{A}(t-1), U_{Y(t)})
\end{aligned}$$

b. Proposed causal question

What is the Causal Question (or questions) of interest for your dataset?

For our project, we plan to ask whether frequency of primary care and hepatology visits at UCSF affect the likelihood of developing hepatocellular carcinoma (HCC) by the end of the follow up period among patients diagnosed with chronic hepatitis C (HCV), who were HCC-free at the beginning of the follow up interval.

What is the ideal experiment that would answer your Causal Question?

The ideal experiment to answer our Causal Question would be to deny people access to primary care and hepatology visits, and see how many developed HCC, then roll back the clock and give them access to just one visit annually (primary care OR hepatology) and see how many developed HCC, then roll back the clock and give them access to 5 visits annually and see how many developed HCC, and so on.

Which of your variables would you intervene on to answer your Causal Question(s)? What values would you set them equal to?

We would intervene on $\bar{A}(t)$ to answer our Causal Question, and set them all equal to zero, then just one of the A timepoints equal to 1 (ultimately we might be interested in the effect of only A(1)=1 compared to only A(2)=1, compared to only A(3)=1, etc.), then each of them equal to 1. We also intervene on C(t) to ensure that all participants have outcome assessed at the end of the interval.

What outcomes are you interested in? Measured when?

We are interested in whether a patient is diagnosed with hepatocellular carcinoma, a liver cancer that commonly develops in people with liver cirrhosis (including as a result of chronic HCV infection) and has a very low survival rate. For our project, anyone diagnosed with HCC anytime before the final timepoint in the dataset (i.e. the date the data were pulled from the electronic medical record) will be counted as having the outcome.

What are your counterfactual outcomes, and how would you explain them in words?

Our counterfactual outcomes are the prevalence of HCC at the end of study follow-up if no one had any primary care or hepatology visits, the prevalence of HCC at the end of study follow-up if everyone had one primary care or hepatology visit over the study period, and the prevalence of HCC at the end of study follow-up if everyone had at least one primary care or hepatology visit over the 5 year study period.

What aspects of the counterfactual outcome distribution are you interested in contrasting?

We are interested in contrasting the counterfactual outcome from various numbers of primary care and/or hepatology visits with the counterfactual outcome from fewer visits, to see if there is some sort of exposure-response relationship.

What contrast are you interested in (e.g., absolute difference? relative difference? MSMs? conditional on subgroups?)?

We are interested in a MSM that helps us understand the relationship between frequency of primary care or hepatology visits and risk of HCC diagnosis, conditional on FIB-4 score and a variety of other demographics.

How would you intervene on the SCM you came up with to evaluate the causal target parameter?

We will intervene to deterministically set $\bar{C}(t) = 1$ and $\bar{A}(t) = \bar{a}(t)$.

c. Our Observed Data

Many of the variables that we intend to use will be coming in the next data pull from the EMR, so we can't present histograms or tables of counts yet. Instead we've listed below each of the variables that we will have after the next data query and their variable types; we listed details for the variables that we *do* have already.

- Number of Visits (annually)- this will be a count variable, at each year
- Diagnosed with HCC (annually)- this will be a binary yes/no
- FIB-4 Score (annually)- Based on previous literature (Sterling et al 2006), we expect these scores to range 0.2 to 10, with much of the probability mass below 1.
- Gender- this will be a categorical variable, likely with three categories (man, woman and non-binary). Based on a prior (small) study of HCV patients at UCSF (Burman et al 2016), it is expected that the cohort will be approximately 70% men.
- Race- this will be a categorical variable. Based on a prior (small) study of HCV patients at UCSF (Burman et al 2016), it is expected that the cohort will be approximately 40% White, 20% Latinx, 30% Black and 10% other races.
- SES- we are using insurance type (MediCal or private) as a marker of SES status, which will be a categorical variable.
- Delta (annually)- is an indicator of missingness for FIB-4
- C (annually)- is an indicator of whether the patient has been censored. Based on our current data, we estimate that NA patients will be censored in the year 2015, NA in the year 2016, NA in the year 2017, and NA in the year 2018.

Missingness

There will be some patients that do not have a FIB-4 score in a given year because they did not have a laboratory assessment of their platelets or transaminases. At each year, the FIB-4 used will be the most recent one calculated and there will also be a variable representing time since FIB-4 was measured.

There is not missingness expected in the exposure-since EMRs are designed for clinical billing, the data on whether or not visit(s) occurred are expected to be highly accurate. Because HCC is a common and severe complication of HCV, we are comfortable assuming that a patient who is still followed in the system and does not yet have a diagnosis of HCC, is truly negative for HCC, rather than simply missing that data.Further, we will intervene at the end to ensure everyone

Patients can be censored from the dataset in one of two ways: either by no longer seeking care within the UCSF system, or if they are deceased.

d. Identification Result and Estimand

Under what assumptions is the target causal parameter identified as a function of the observed data distribution?

Our target causal parameter is identified as a function of the observed data distribution under the assumptions that there is independence between each of the exogenous variables (i.e. there are no shared unknown variables influencing each of the endogenous nodes in our SCM) and that there are no practical positivity violations (i.e. there is a >0 probability of HCC among all of the baseline covariate and treatment regime combinations). In addition to controlling for any baseline covariates (Ws), we also must rely on the sequential randomization assumption for our data generating process.

What is your $\Psi(P_0)$, the statistical estimand?

*Our statistical estimand is:

$$\Psi(P_0) = m(\bar{a}|\beta) = E[Y_{\bar{a}}] = \beta_0 + \beta_1 \sum_{t=1}^{8} a(t)$$

3. Estimation

a. What estimators will you use?

b. How will you implement these estimators?

4. Preliminary Results

a. Simulation

To run our simulation, we first create a dataframe O which includes all of the exogenous and endogenous variables in our SCM. Each of our $U_W, U_{C(t)}$, and $U_{Y(t)}$ variables have a uniform distribution with a min of 0 and max of 1. Each of our $U_{L(t)}$ variables have a gamma distribution with a shape parameter of 1.5 and scale parameter of 2. Each of our $U_{A(t)}$ variables have a poisson distribution with a rate parameter (λ) of 3.

For our endogenous variables:

- For W we created a nominal categorical variable for the 16 possible different combinations of race, gender, and SES that apply to our dataset.
- For the $L(t)$ variables, we set L(1) equal to the underlying $U_{L(1)}$ gamma distribution and a linear combination of years since FIB-4 was measured. In subsequent years, we generated a variable for whether a new FIB-4 was measured that year. If so, we set $L(t)$ to the FIB-4 score from the prior year ($L(t-1)$) and add 10% of the value generated by the underlying gamma distribution for $U_{L(t)}$ - i.e. we expect FIB-4 score to increase slightly each year as people's liver disease slowly progresses. If FIB-4 score was not measured in subsequent years then we set FIB-4 equal to the prior FIB-4 score, with a variable for how many years since it was measured.
- For $A(1)$, if there is no FIB-4 score known for that timepoint, we use the underlying poisson distribution for $U_{A(t)}$, plus two extra visits for people diagnosed with HCC at that timepoint ($Y(t) = 1$). When the FIB-4 score is known ($\Delta(t) = 0$) then $A(t)$ is calculated by using the underlying poisson distribution for $U_{A(t)}$ plus FIB-4 score (adding more visits for higher FIB-4 scores, indicating worsening cirrhosis) plus two extra visits for people diagnosed with HCC at that timepoint ($Y(t) = 1$).
- For the $C(t)$ variables, if a subject had no primary care or hepatology visits in the prior year ($A(t-1) = 0$) we set $C(t)$ to a 80% chance of being censored; if they had at least one visit in the prior year then we set $C(t)$ to only an 8% chance of being censored.
- For $Y(t)$ we calculated it as having a 10% chance of indicating HCC diagnosis if they had no visits over the interval, with a decreasing chance of HCC with every visit the subject had in the prior year and an increasing chance of HCC as the FIB-4 score rises, indicating worsening cirrhosis ($Y(t) = I(U_{Y(t)} + 0.01(A(t-1)) - 0.05(L(t-1))) < 0.03$).

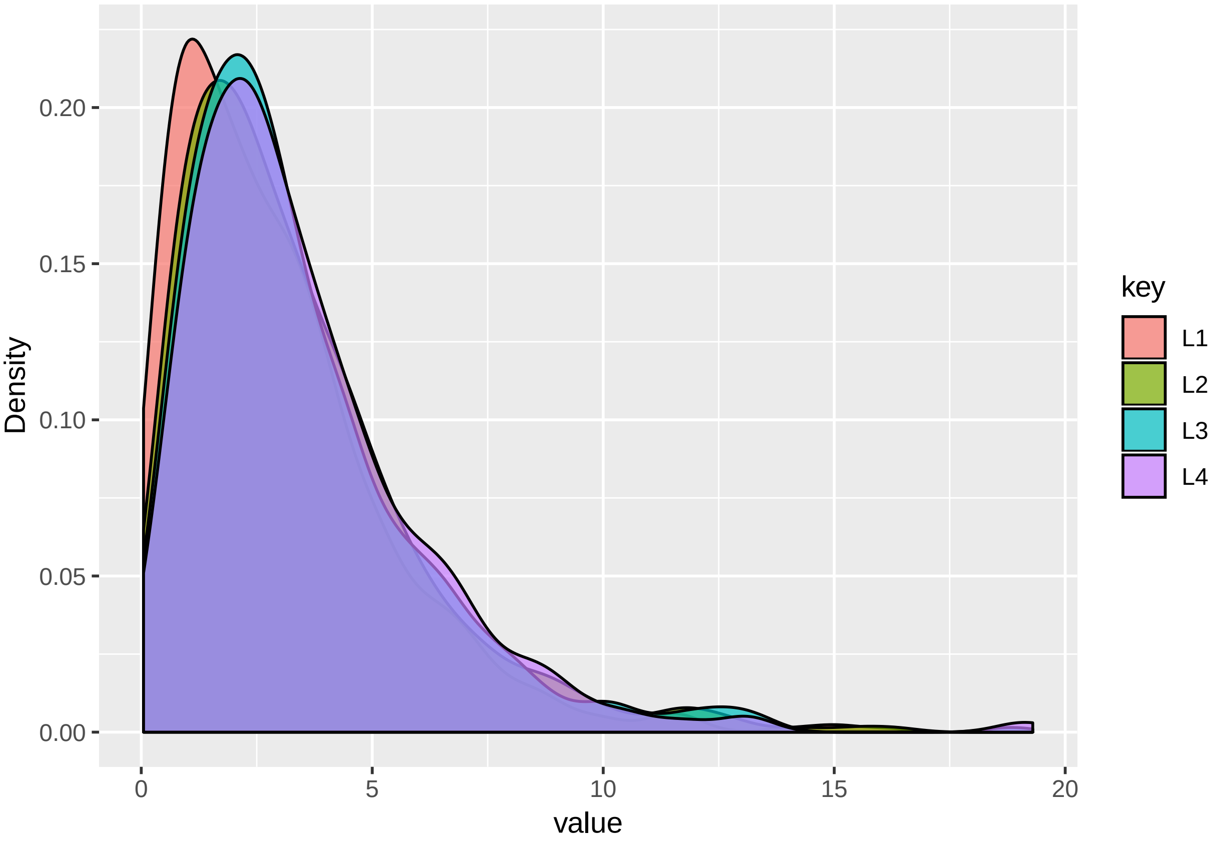
We then set a seed so our numbers could be replicated exactly, and generated data with n = 1000. The following table and plots describe the results of the simulation.

Cases of HCC in the Simulated Data

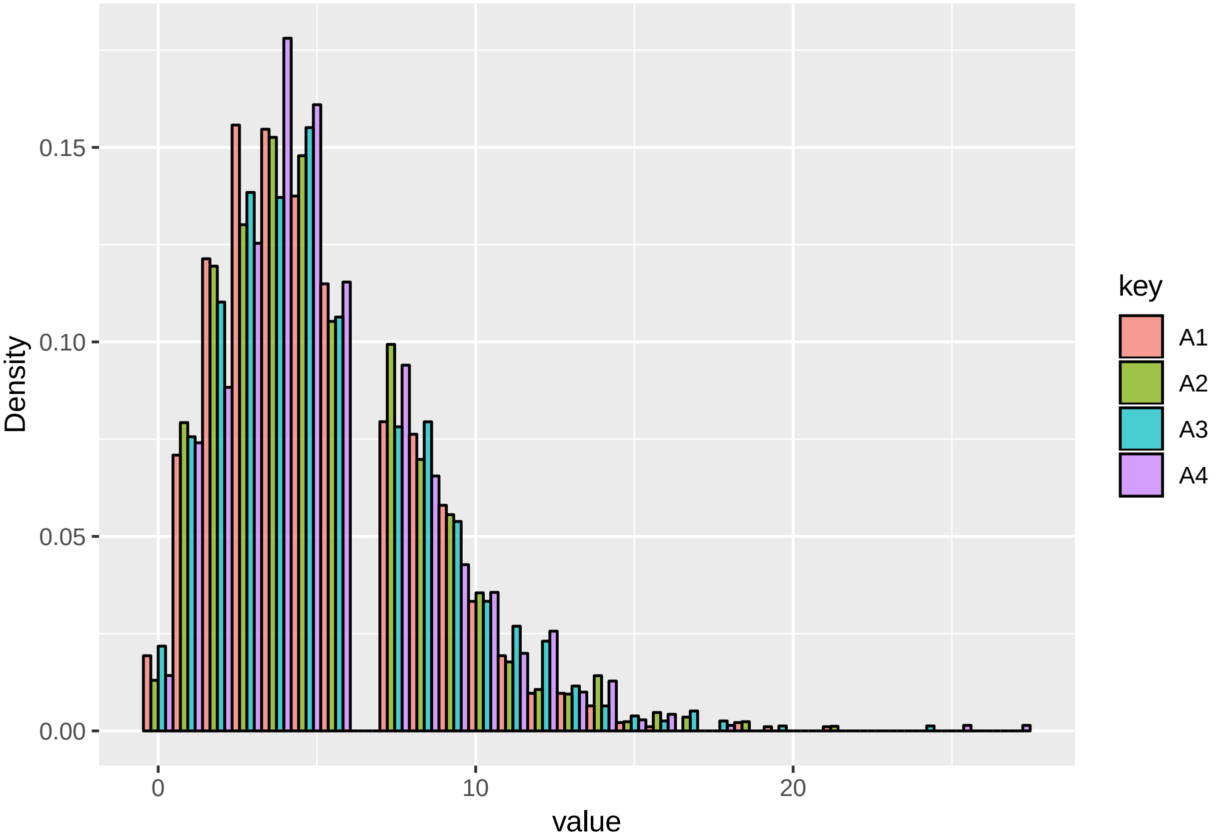
In our simulated dataset, there were 0 cases of hepatocellular carcinoma (HCC) by the end of the study interval.

Histograms of Visits and FIB-4 Scores in the Simulated Data

Here are density plots demonstrating the distribution of FIB-4 scores over the years they were measured.



Here are histograms of the number of visits each patient had per year.



Implement our intervention computationally.

To do this, we use the simulation code written earlier, but adapt our function, replacing the structural equations for $\bar{C}(t)$ and $\bar{A}(t)$ with flexible parameters allowing us to specify values for our intervention.

Generate many counterfactual outcomes, then evaluate $\Psi^F(P_{U,X})$

For simplicity, we simulated counterfactual outcomes to try to determine $\Psi^F(P_{U,X})$ as the ratio between the proportion of patients diagnosed with HCC under the counterfactual scenario where $\bar{A}(t) = 6$ (all people had 6 visits per year, an extremely high number of visits), and $\bar{A} = 0$ (all people had no primary care or hepatology visits in any year). If we simulate data with n = 1000, then* $\Psi^F(P_{U,X}) = 1.302$.

Write a sentence interpreting your $\Psi^F(P_{U,X})$

This means that people with very frequent primary care or hepatology visits ($\bar{A} = 6$ used to represent that in this example) are 30.2% more likely to develop HCC than people who have no primary care or hepatology visits ($\bar{A} = 0$) in the five years under study.

b. Data example

5. References

Burman BE, Bacchetti P, Khalili M. Moderate Alcohol Use and Insulin Action in Chronic Hepatitis C Infection. Dig Dis Sci. 2016;61(8):2417-2425. doi:10.1007/s10620-016-4119-0

Sterling, R. K., Lissen, E. , Clumeck, N. , Sola, R. , Correa, M. C., Montaner, J. , S. Sulkowski, M. , Torriani, F. J., Dieterich, D. T., Thomas, D. L., Messinger, D. and Nelson, M. (2006), Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology, 43: 1317-1325. doi:10.1002/hep.21178