252E Final Project Part 1: Relationship between Frequency of Care and Progression to Hepatocellular Carcinoma among Chronic Hepatitis C Patients

Stephanie Holm and Shelley Facente

10/3/2019

1. Description of our Dataset

We have access to a dataset of chronic Hepatitis C patients currently receiving care in the UCSF system. There are 1937 patients, who are seen in a variety of primary care clinics and the hepatology (liver) clinic. These data come from a query of Apex, the UCSF-specific build of the electronic medical record system Epic. We have a preliminary dataset resulting from the intial Apex query and will be getting access to more data soon. This initial report has been completed with our intended causal question, describing the data that we currently have and indicating where we are waiting for more data.

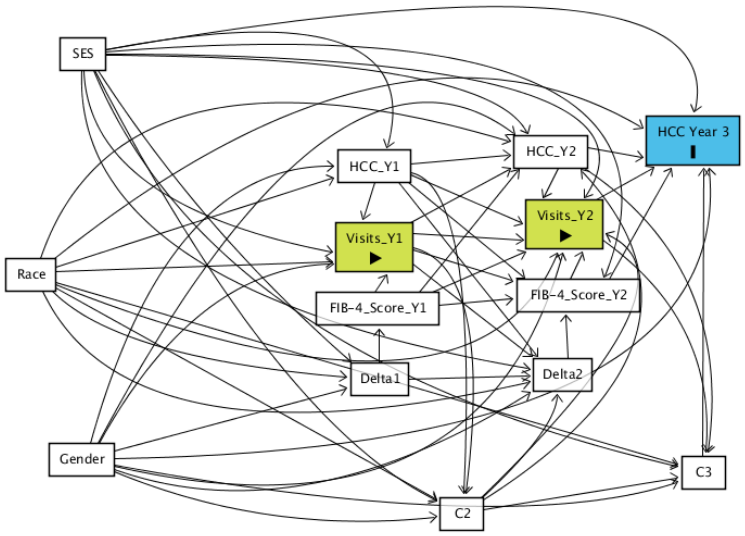
A. Defining Our Variables

Exposures: We will have data on the annual number of clinical visits that each chronic HCV patient had in the UCSF system over the last five years.

Outcome: The outcome will be diagnosis of hepatocellular carcinoma (HCC), which occured in 440 patients. We do not currently have the dates of the HCC diagnoses for all patients, however, for the subset that had biopsies (n =63) 84.13% of them occured after 01/01/2015 suggesting that the large majority of the hepatocellular carcinoma diagnoses occured in the last 5 years.

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

B. Our DAG



We intend to do this analysis using 5 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 delta

variable indicating whether or not our covariate was measured and a C value at each time point indicating whether or not the patient has been censored from the dataset.

C. 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, gender and SES)
- C(t) is an indicator of being censored at time t (1 means they were censored)
- $\Delta(t)$ is an indicator of having missing covariate data at time t (1 indicates missing)
- L(t) is the covariate (FIB-4 score) at time t
- A(t) is the exposure (number of visits)
- Y(t) is the outcome (an indicator of HCC diagnosis)

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

Structural Equations, F:

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

D. Exploring Our 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. Below that we've presented the distribution of some of the related 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 102 patients will be censored in the year 2015, 186 in the year 2016, 230 in the year 2017, and 327 in the year 2018.

E. 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. We anticipate that missingness in FIB-4 may be related to baseline demographic factors, as well as the number of visits at the prior time point and prior diagnosis of HCC.

There is not missingness expected in the exposure-since EMRs are designed for clinical billing, the data on whether or not visit(s) occured 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.

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.

F. 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 $U_W, U_{C(t)}, U_{Y(t)},$ and $U_{\Delta(t)}$ 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 $U_{L(t)}$ variables have a poisson distribution with a rate parameter (λ) of 3. For our endogenous variable W we create a nominal categorical variable for the 16 possible different combinations of race, gender, and SES that apply to our dataset. For the $\Delta(t)$ variables, if $U_{\Delta(t)}$

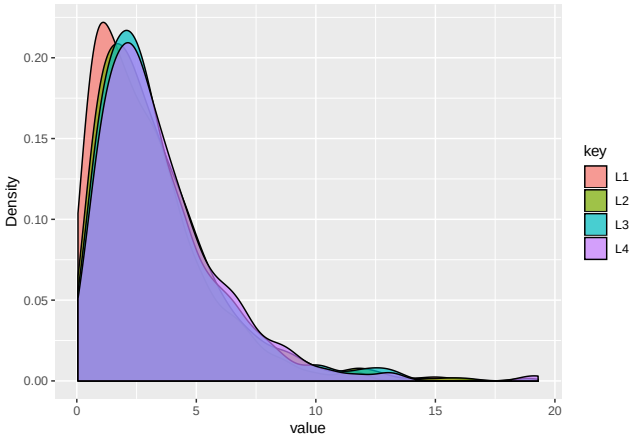
We then set a seed so our numbers could be replicated exactly, and generated data with n = 1000.

Table of Cases of HCC and missing data by Year in the Simulated Data

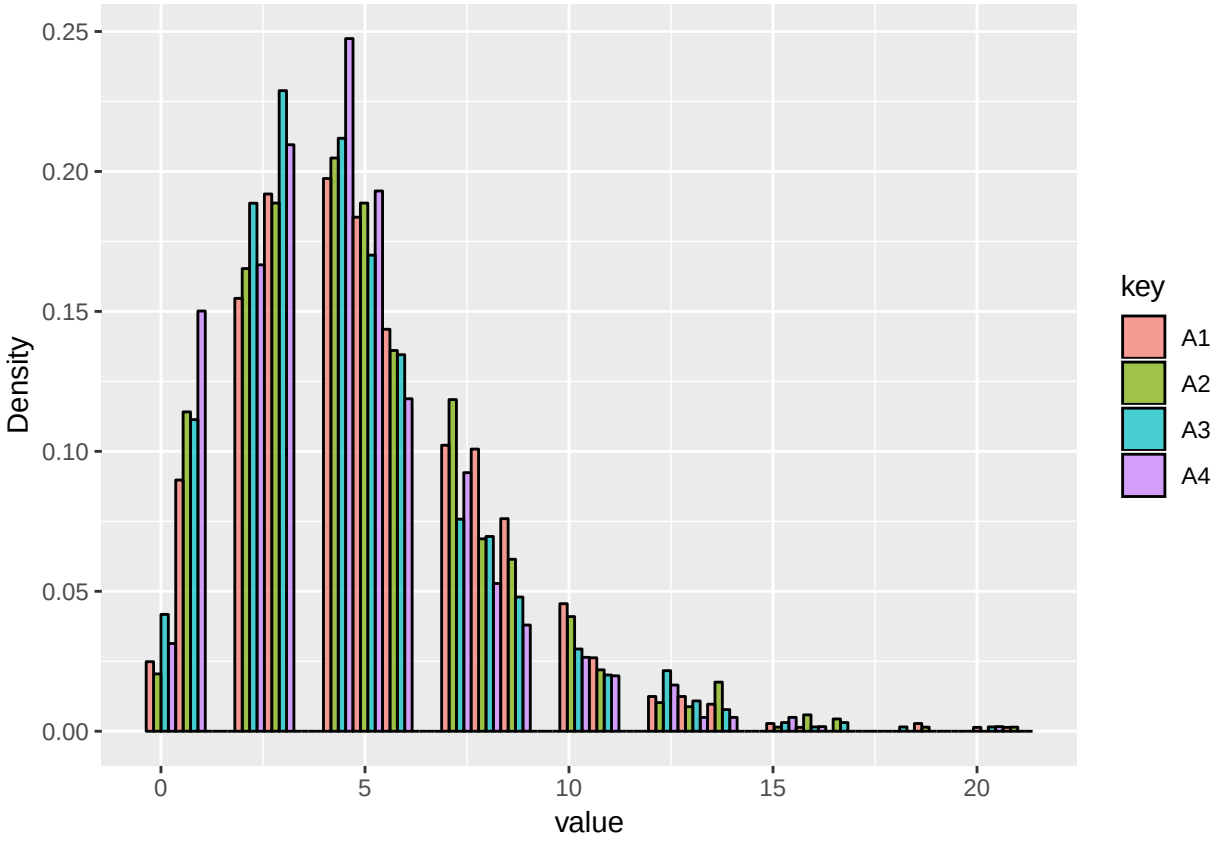
	Year 1	Year 2	Year 3	Year 4	Year 5
Cases of HCC	36	106	159	203	234
Patients Censored	0	92	162	246	312
Patients Missing FIB-4	304	440	505	577	NA

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

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



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



2. 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) among patients diagnosed with chronic hepatitis C (HCV).

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 (primary care OR hepatology) in 5 years and see how many developed HCC, then roll back the clock and give them access to 5 visits in 5 years 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.

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.

Target parameter and counterfactual outcomes

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 5 year 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.

Come up with a target parameter that would answer your Causal Question.

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

We are interested in contrasting the counterfactual outcome from a large number of primary care and/or hepatology visits with the counterfactual outcome from few or no visits, to see if there is some sort of dose-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.

Intervention on the SCM

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

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

Implement this intervention computationally.

First we make a matrix of every possible treatment regime permutation, then use a "for" loop to generate counterfactual outcomes for each regime. Then we use this in a logistic regression using ordinary least squares, with $beta_1$ as the target parameter.

Evaluate $\Psi^F(P_{U,X})$

WILL NEED TO FILL THIS IN

Using simulation, generate many counterfactual outcomes, then evaluate $\Psi^F(P_{U,X})$.

If we use the data we simulated with n = 1000,

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

3. Identification and Estimand

Under what assumptions is the target causal parameter you came up with in the previous lab identified as a function of the observed data distribution?

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

Confirm that in your simulation, the value of your estimand equals the value of your target causal parameter.

4. Preliminary Feasibility assessment

# of observations (individuals) meeting inclusion criteria:

# of individuals following each regime of interest:

Basic descriptive stats for outcome (eg if binary, how many events):

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