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

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.

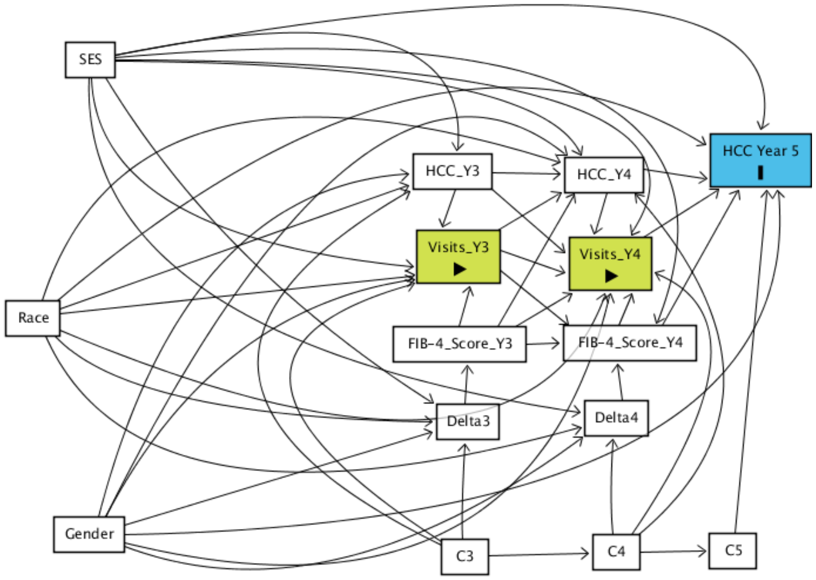
Defining Our Variables

-**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.)

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.

Our Structural Causal Model

-Using those relationships drawn in the previous step, define your structural equations generically; in other words, don't assuming distributions or functional forms yet. Do you have any prior knowledge on the functional forms?

Exploring Our Data

-Make histograms for continuous variables and tables for binary/categorical variables. -What shapes do the distributions seem to take? Based on the shape, what known distribution do you think that variable's error term is drawn from?

-If you've picked a distribution for an exogenous variable, how would you parameterize it?

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. There is not missingness 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 fo 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.

-Based on *a priori* knowledge, do you expect missingness to depend on other variables? Which ones?

Simulation

-Come up with more specific structural equations that relate the endogenous variables based on the previous two questions.

-Create a function to simulate your data and generate $n = 1000$ copies of your O.

-Check the histograms and summary statistics of the variables in your simulated data and see how they match up to your real data. Over the course of the class you can refine your data generating structure to match the data at hand.

Proposed Causal Question

-Inclusion criteria

-Outcome

-Intervention Nodes

-Counterfactual outcomes of interest

-Target Causal parameter

Defining your Causal Question}

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

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

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

-What outcomes are you interested in? Measured when?

-Target parameter and counterfactual outcomes

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

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

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

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

Intervention on SCM

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

-Implement this intervention computationally.

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

-Using simulations, generate many counterfactual outcomes.

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

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

Identification and Estimand

1. Under what assumptions is the target causal parameter you came up with in the previous lab identified as afunction of the observed data distribution?
2. What is your $\psi(P_0)$, the statistical estimand?
3. Optional: confirm that in your simulation, the value of your estimand equals the value of your target causalparameter.

Preliminary Feasibility assessment