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.

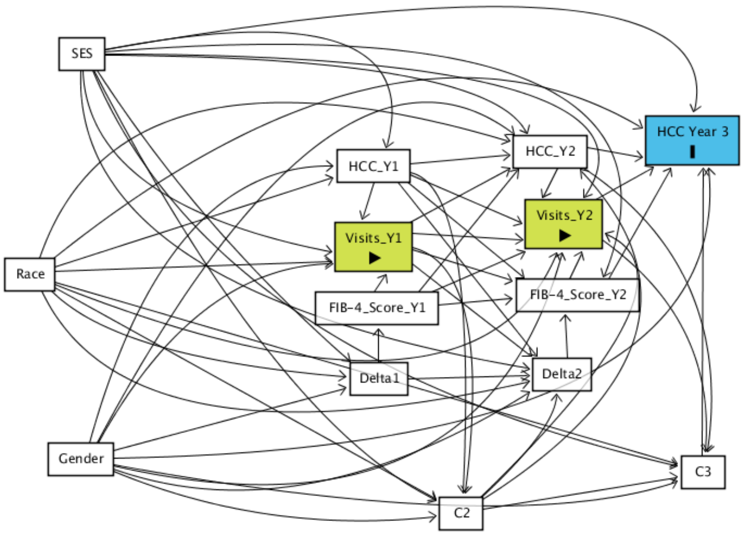
A. Defining Our Variables

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

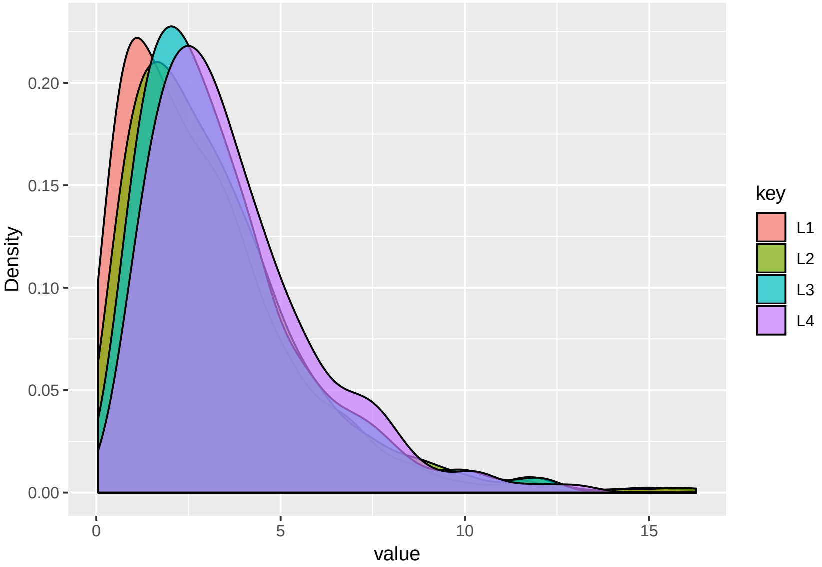
```
generate_data <- function(n) {
  U <- data.frame(UW = runif(n = n, min = 0, max = 1),
                  UD1 = runif(n = n, min = 0, max = 1),
                  UL1 = rgamma(n = n, shape = 1.5, scale = 2),
                  UY1 = runif(n = n, min = 0, max = 1),
                  UA1 = rpois(n = n, lambda = 3),
                  UC2 = runif(n = n, min = 0, max = 1),
                  UD2 = runif(n = n, min = 0, max = 1),
                  UL2 = rgamma(n = n, shape = 1.5, scale = 2),
                  UY2 = runif(n = n, min = 0, max = 1),
                  UA2 = rpois(n = n, lambda = 3),
                  UC3 = runif(n = n, min = 0, max = 1),
                  UD3 = runif(n = n, min = 0, max = 1),
                  UL3 = rgamma(n = n, shape = 1.5, scale = 2),
                  UY3 = runif(n = n, min = 0, max = 1),
                  UA3 = rpois(n = n, lambda = 3),
                  UC4 = runif(n = n, min = 0, max = 1),
                  UD4 = runif(n = n, min = 0, max = 1),
                  UL4 = rgamma(n = n, shape = 1.5, scale = 2),
                  UY4 = runif(n = n, min = 0, max = 1),
                  UA4 = rpois(n = n, lambda = 3),
                  UC5 = runif(n = n, min = 0, max = 1),
                  UY5 = runif(n = n, min = 0, max = 1))

  O <- U
  O <- O %>%
    mutate(W = ifelse(UW < 1/16, 1,
                      ifelse(UW<2/16, 2,
                             ifelse(UW<3/16, 3,
                                    ifelse(UW<4/16,4,
                                           ifelse(UW<5/16, 5,
                                                  ifelse(UW<6/16, 6,
                                                         ifelse(UW<7/16, 7,
                        ifelse(UW<8/16, 8,
                               ifelse(UW<9/16, 9,
                                      ifelse(UW<10/16, 10,
                                             ifelse(UW<11/16, 11,
                                                    ifelse(UW<12/16, 12,
                                                           ifelse(UW<13/16, 13,
                                                                  ifelse(UW<14/16, 14,
                        ifelse(UW<15/16, 15, 16))))))))))))))),
           D1 = as.numeric(UD1<0.3),
           L1 = ifelse(D1 == 0, UL1, NA),
           Y1 = as.numeric(UY1 < 0.06),
           A1 = ifelse(is.na(L1), round(as.numeric(UA1 + 2*Y1)),
                                  round(as.numeric(UA1 + L1 + 2*Y1))),
           C2 = ifelse(A1 == 0, as.numeric(UC2<0.6),
                                               as.numeric(UC2<0.05)),
           D2 = ifelse(C2 == 1, 1,
                                ifelse(D1 == 1, as.numeric(UD2<0.6),
                                                as.numeric(UD2<0.3))),
           L2 = ifelse(D2 == 0, L1 + 0.1*UL2, NA),
           Y2 = ifelse(Y1 == 1, 1,
                                ifelse(C2 == 0,
                                       ifelse(is.na(L1),
                                              as.numeric((UY2+0.01*A1)< 0.06),
                                              as.numeric((UY2+0.01*A1-0.05*L1)< 0.06)),
                                                NA)),
           A2 = ifelse(C2 == 0, ifelse(is.na(L2), round(as.numeric(UA2 + 2*Y2)),
                                                  round(as.numeric(UA2 + L2+ 2*Y2))),
                                       NA),
           C3 = ifelse(C2 ==1, 1,
                               ifelse(A2 == 0, as.numeric(UC3<0.6),
                                               as.numeric(UC3<0.05))),
           D3 = ifelse(C3 == 1, 1,
                                ifelse(D2 == 1, as.numeric(UD3<0.6),
                                                as.numeric(UD3<0.3))),
           L3 = ifelse(D3 == 0, L2 + 0.1*UL3, NA),
           Y3 = ifelse(Y2 == 1, 1,
                                ifelse(C3 == 0,
                                       ifelse(is.na(L2),
                                              as.numeric((UY3+0.01*A2)< 0.06),
                                              as.numeric((UY3+0.01*A2-0.05*L2)< 0.06)),
                                                NA)),
           A3 = ifelse(C3 == 0, ifelse(is.na(L3), round(as.numeric(UA3 + 2*Y3)),
                                                  round(as.numeric(UA3 + L3+ 2*Y3))),
                                NA),
           C4 = ifelse(C3 == 1, 1,
                               ifelse(A3 == 0, as.numeric(UC4<0.6),
                                               as.numeric(UC4<0.05))),
           D4 = ifelse(C4 == 1, 1,
                                ifelse(D3 == 1, as.numeric(UD4<0.6),
                                                as.numeric(UD4<0.3))),
           L4 = ifelse(D4 == 0, L3 + 0.1*UL4, NA),
           Y4 = ifelse(Y3 == 1, 1,
                                ifelse(C4 == 0,
                                       ifelse(is.na(L3),
                                              as.numeric((UY4+0.01*A3)< 0.06),
                                              as.numeric((UY4+0.01*A3-0.05*L3)< 0.06)),
                                                NA)),
           A4 = ifelse(C4 == 0, ifelse(is.na(L4), round(as.numeric(UA4 + 2*Y4)),
                                                  round(as.numeric(UA4 + L4+ 2*Y4))),
                                NA),
           C5 = ifelse(C4 ==1, 1,
                               ifelse(A4 == 0, as.numeric(UC5<0.6),
                                               as.numeric(UC5<0.05))),
           Y5 = ifelse(Y4 == 1, 1,
                                ifelse(C5 == 0,
                                       ifelse(is.na(L4),
                                              as.numeric((UY5+0.01*A4)< 0.06),
                                              as.numeric((UY5+0.01*A4-0.05*L4)< 0.06)),
                                                NA)))
  return(0)
}

set.seed(8675309)
SimData <- generate_data(1000)
```

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

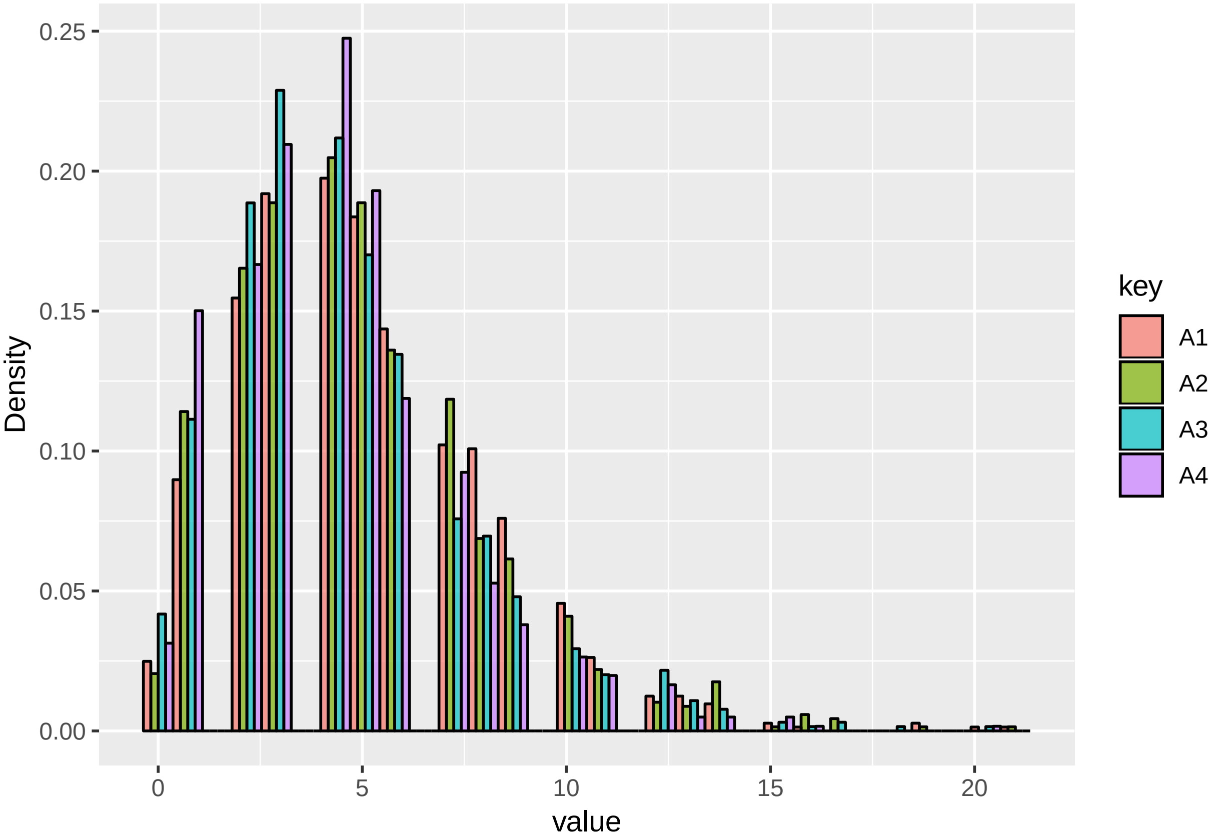
	Year 1	Year 2	Year 3	Year 4	Year 5
Cases of HCC	62	156	218	256	287
Patients Censored	0	56	107	163	211
Patients Missing FIB-4	304	521	688	797	NA

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?

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

3. Identification and Estimand

1. 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?
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.

4. Preliminary Feasibility assessment

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