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

11/23/2019

1. Background

a. Brief Introduction

Hepatocellular carcinoma (HCC), a type of liver cancer, is the second most common cause of liver-related death throughout the world (Bosetti et al 2014). HCC is a major complication of infection with hepatitis C (HCV) virus, occuring in 1%-4% of people with liver cirrhosis each year (Omland et al 2010). HCV becomes chronic infection for about 80% of adults, and if left untreated can cause increasing cirrhosis over a period of 20-30 years of often asymptomatic disease, until damage is severe and major complications are irreversable (El-Serag 2012). Certain HCV viral genotypes (particularly genotype 3) are associated with higher risk of HCC, as is continued smoking (Chuang et al 2010) and alcohol use (Donato et al 2002), as well as concurrent diabetes or obesity (Huang et al 2017, Calle et al 2003).

Since 2014, direct-acting antiviral (DAA) therapy has dramatically improved prognosis for people living with HCV; 8-12 weeks of well-tolerated oral therapy leads to cure in more than 90% of patients (Burstow et al 2017), halting cirrhosis progression and apparently reducing the elevated risk of HCC (Axley et al 2017). Patients whose HCV has already caused some cirrhosis continue to be at some increased risk of HCC even post-cure, and given the high mortality risk it is critical that people with HCV infection receive systematic monitoring with liver ultrasound (the European Association for Study of the Liver currently recommends screening twice per year) (EASL 2015). Therefore, we hypothesize that regular visits to a primary care physician or hepatologist who can provide screening and ongoing disease management (including curative treatment for HCV infection in most cases) is preventive for development of HCC among people with known chronic HCV infection.

b. Description of our Dataset

We are using a dataset of chronic hepatitis C patients receiving care in the UCSF system since 2009. There are 2297 patients, who were adults by the start of 2015 and have been seen in a variety of primary care clinics and the hepatology (liver) clinic between 2015 and 2019. These data come from a query of Apex, the UCSF-specific build of the electronic medical record system Epic.

2. Roadmap

a. Causal Model

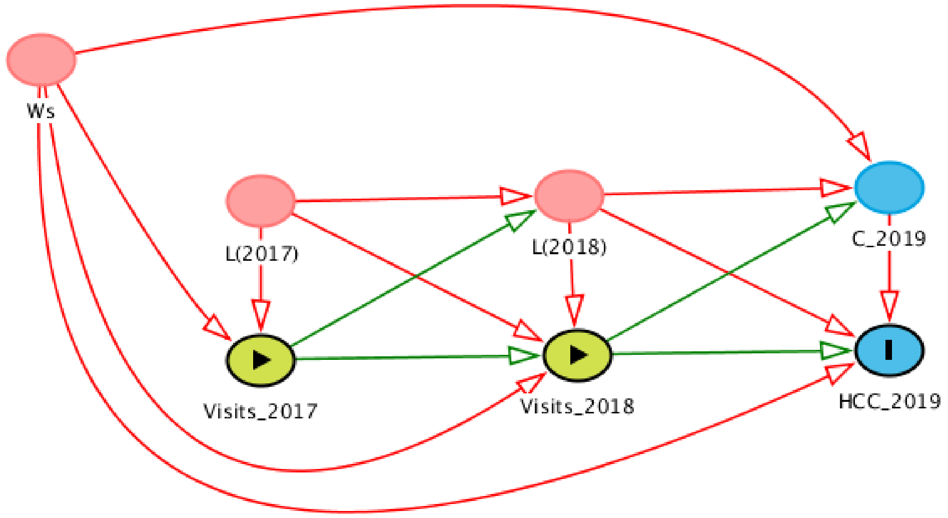
i. Defining Our Variables

Exposures: We have data on the annual number of clinical visits that each chronic HCV patient had in the UCSF system annually from 2015 through 2018.

Outcome: The outcome is diagnosis of HCC, by the final year (2019) which occured in 397 patients.

Covariates: FIB-4 score (a validated measure used for prediction of cirrhosis (Khan et al 2017), which uses age, platelet count and liver transaminases), years since FIB-4 score last measured, biologic sex, race, insurance type (Medi-Cal vs private- a surrogate of socioeconomic status.)

ii. Our DAG



This analysis includes 5 years worth of data (4 of exposure and 1 of outcome), but in order to simplify the DAG we present only the final 3 years (two of exposure data and the final year of outcome) here. This DAG includes a C value indicating whether or not the patient had the outcome measured by the end of the study.

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

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(K) 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, $\mathcal{M}^{\mathcal{F}}$, for t from 1 to 5:

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

We are also operating with one key exclusion restriction: once someone develops the outcome $[Y(t) = 1]$ then we set their $A(t+1...K)$ and $L(t+1, ..., K)$ to NA for the remainder of the analysis, and set $C(K) = 1$ for that subject.

b. Proposed causal question

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

For our project, we 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 1 additional visit annually and see how many developed HCC, then roll back the clock and give them access to 1 additional visit 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(K) 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 HCC, measured at each timepoint (by the end of each year in our study period). For our time-to-event analysis, 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 during the period of study.

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

Our counterfactual outcomes are the prevalence of HCC by the end of study follow-up if no one had any primary care or hepatology visits, and the difference in prevalence of HCC at the end of study follow-up for every additional annual primary care or hepatology visit (up to a maximum of four annually) over the study period (i.e. the association between the number of primary care or hepatology visits each year and the hazard of developing HCC.

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 (as a proxy for liver cirrhosis) and a variety of other demographics.

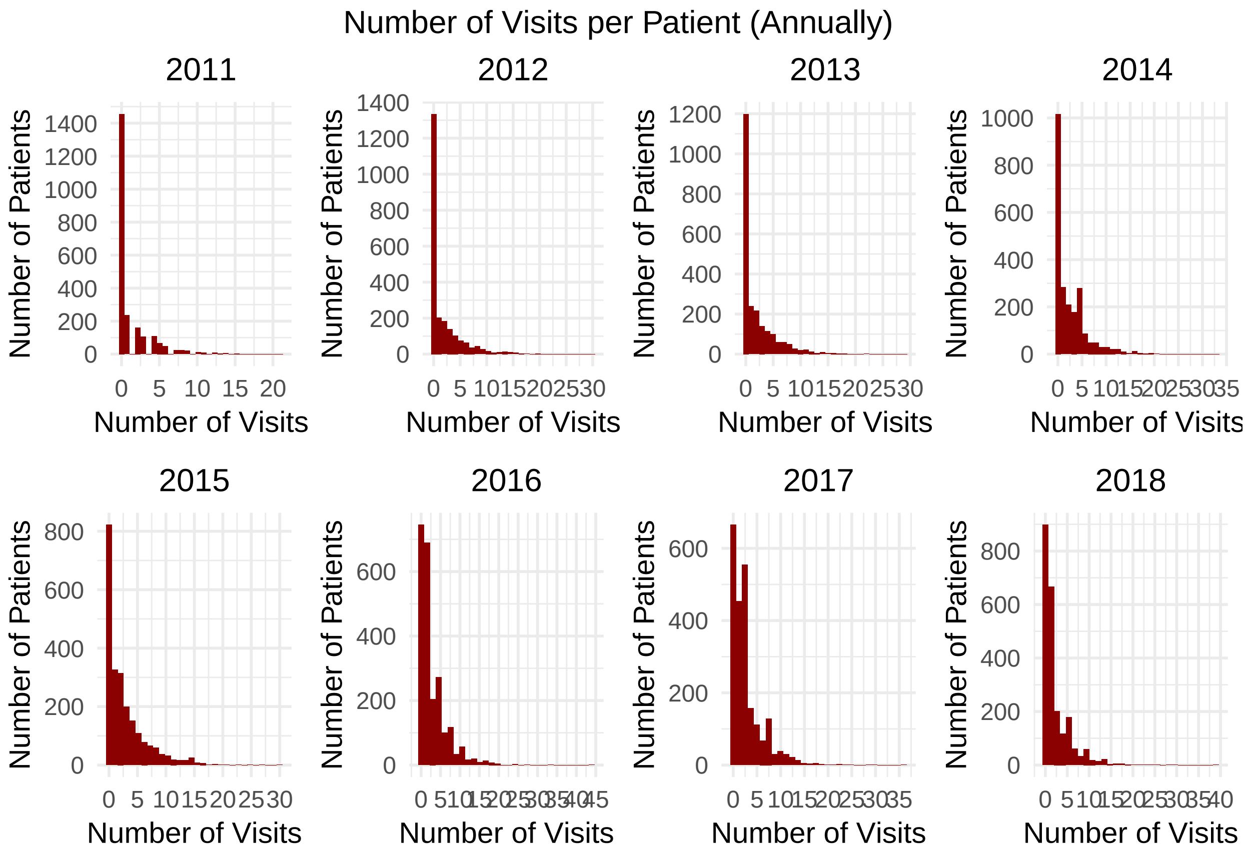
How would you intervene on the SCM you came up with to evaluate the causal target parameter?

We will intervene to deterministically set $\bar{C}(K) = 0$ and $\bar{A}(t) = \bar{a}(t)$.

c. Our Observed Data

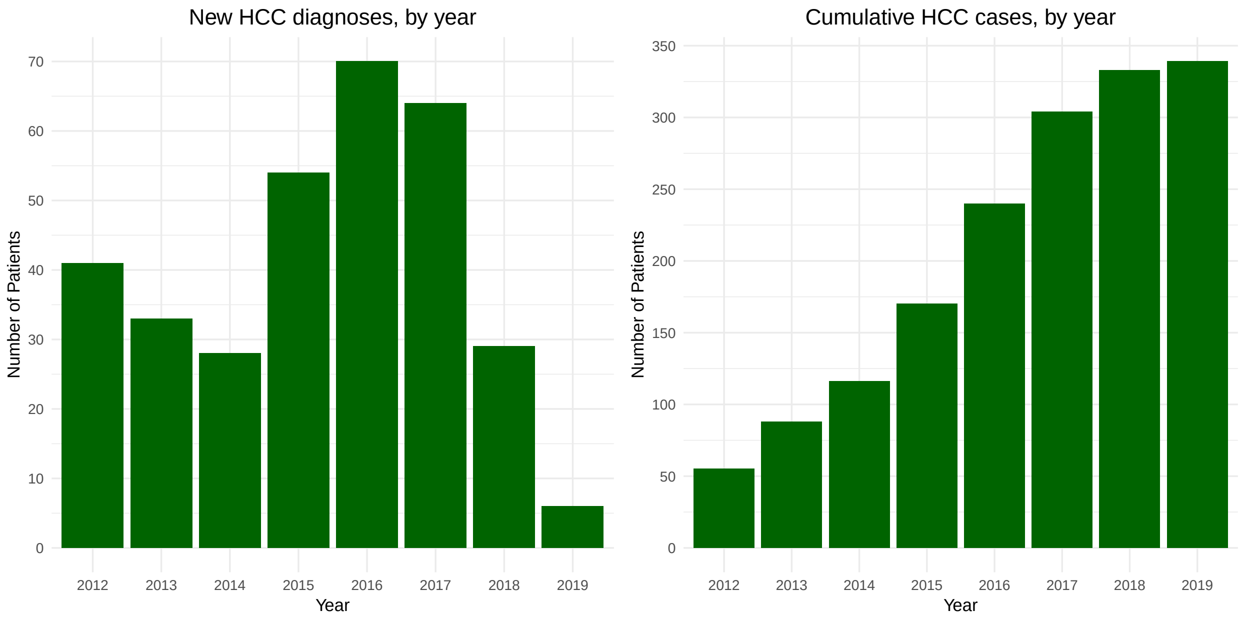
Number of Visits (annually)

The number of visits per patient ranges from 0 to 44 in any given year, with a median of 1 visit per year overall. The figure below presents number of visits per patient in each year, truncated at 25 visits.



Diagnosed with HCC (annually)

397 people (17.3%) developed the outcome (were diagnosed with HCC during the course of the study). The figure below shows the number of people diagnosed with HCC in each year, and cumulatively, throughout the study period (**note** that 2019 is an incomplete year!)

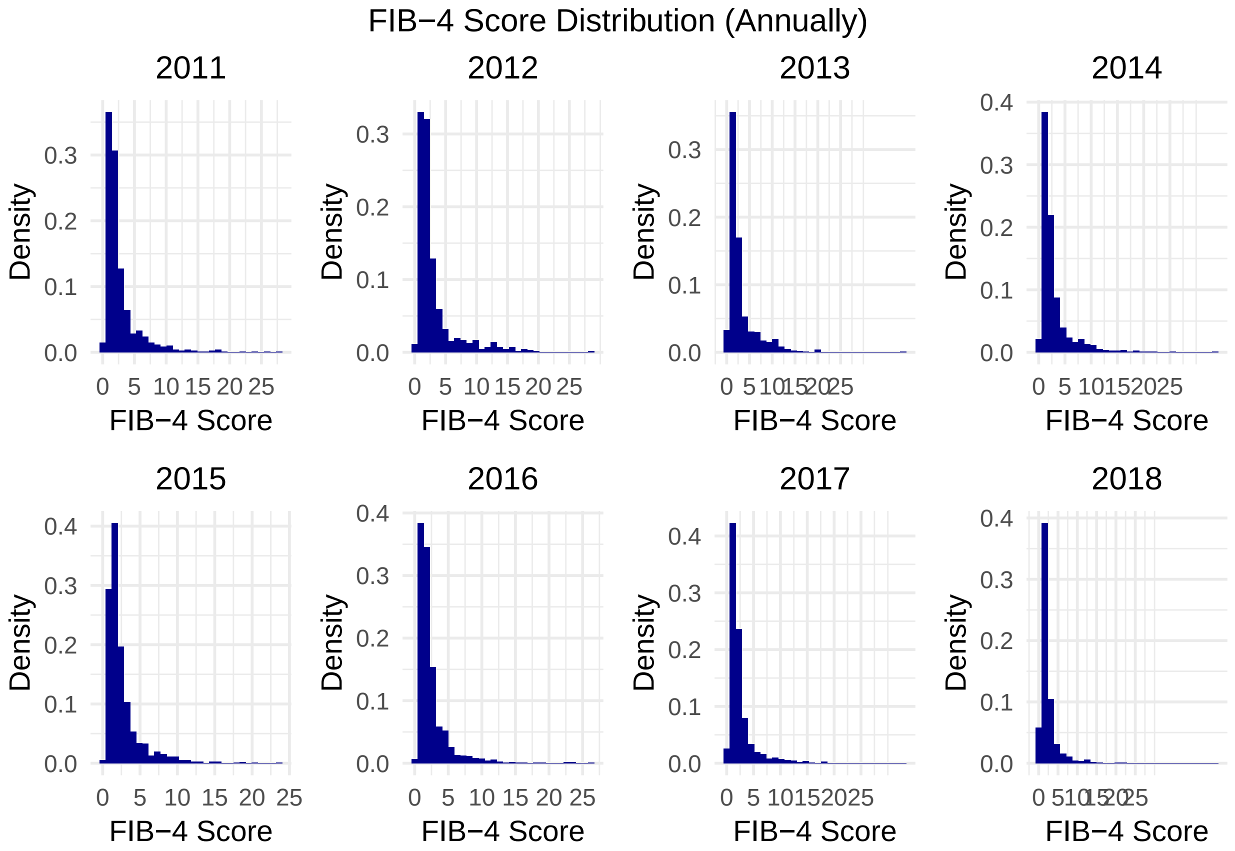


FIB-4 Score (annually)

FIB-4 scores are calculated with inputs of age, platelet count, AST, and ALT. Scores of <1.45 are considered to be strongly suggestive of no liver fibrosis, and scores >3.25 are indicative of advanced fibrosis and/or cirrhosis.

FIB-4 scores in this dataset range of 0.181 to 45.956 , with an IQR of 1.198-2.872 (i.e. the FIB-4 score of 45 is a substantial outlier). 1752 FIB-4 scores are >3.25 throughout all years of follow-up on all patients, indicating likely cirrhosis and increased risk of HCC at those timepoints. There are 390 instances where a patient's FIB-4 score was calculated to be greater than 9.

The figure below presents the distribution of FIB-4 scores in each year.



Other demographics

- **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.

Missingness

Note that we have FIB-4 scores calculated annually (when the appropriate components are available) from 2009 forward. Thus, we have a FIB-4 score available for NA % of participants at the start of the study interval (calculated somewhere in 2009-2015). For the rest of the participants, we used multiple imputation to impute a 2015 FIB-4 score. For all participants, the FIB-4 score and years since measured are updated every year that all the requisite labs were measured. Thus, by the end of interval, NA % of participants had a measured FIB-4 score rather than an imputed value.

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 has outcome measured.

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?

We will estimate our target causal parameter using MSMs estimated with G-computation, IPTW, and LTMLE estimators.

b. How will you implement these estimators?

We will implement each of these estimators using the `ltmleMSM function in the` *`ltmle`* `R package.`

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 0.6 and scale parameter of 2.6. Each of our $U_{A(t)}$ variables have a negative binomial distribution with a dispersion parameter of 4 and a probability of 0.6. The shape and scale parameters for these distributions were chosen by having them reflect the distribution of variables in our observed data.

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. In subsequent years, 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.
- $A(t)$ 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)
- For the $C(t)$ variable, if a subject had no primary care or hepatology visits in the prior year they were censored; if they had at least one visit in the prior year then they were not censored.
- For $Y(t)$ we calculated it as having a 10% chance of indicating HCC diagnosis if they had no visits over the prior 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.

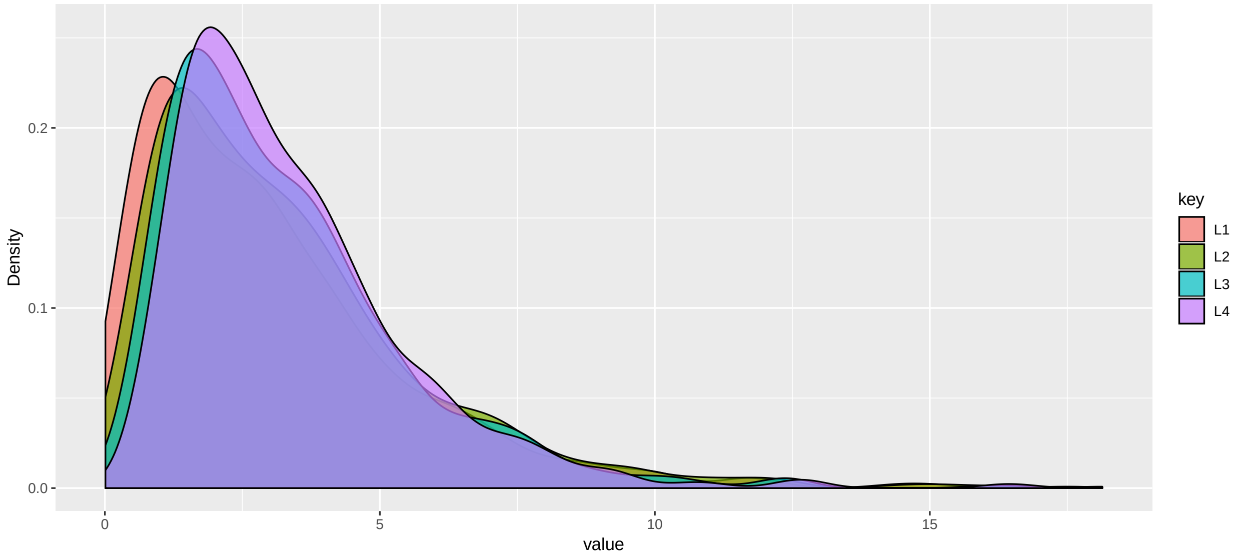
```
generate_data <- function(n) {
  U <- data.frame(UW = runif(n = n, min = 0, max = 1),
                  UL1 = rgamma(n = n, shape = 1.5, scale = 2),
                  UA1 = rnbinom(n=n, size = 4, prob = 0.6),
                  UL2 = rgamma(n = n, shape = 1.5, scale = 2),
                  UA2 = rnbinom(n=n, size = 4, prob = 0.6),
                  UY2 = runif(n = n, min = 0, max = 1),
                  UL3 = rgamma(n = n, shape = 1.5, scale = 2),
                  UA3 = rnbinom(n=n, size = 4, prob = 0.6),
                  UY3 = runif(n = n, min = 0, max = 1),
                  UL4 = rgamma(n = n, shape = 1.5, scale = 2),
                  UA4 = rnbinom(n=n, size = 4, prob = 0.6),
                  UY4 = runif(n = n, min = 0, max = 1),
                  UA5 = rnbinom(n=n, size = 4, prob = 0.6),
                  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))))))))))))))),
           L1 = UL1,
           A1 = round(as.numeric(UA1 + 0.1*L1),0),
           L2 =  L1 + 0.1*UL2,
           A2 = round(as.numeric(UA2 + 0.1*L2),0),
           Y2 = as.numeric((UY2+0.01*A1-0.05*L1)< 0.03),
           L3 = ifelse(Y2==1, NA, L2 + 0.1*UL3),
           A3 = ifelse(Y2==1, NA, round(as.numeric(UA3 + 0.1*L3),0)),
           Y3 = ifelse(Y2==1, 1, as.numeric((UY3+0.01*A2-0.05*L2)< 0.03)),
           L4 = ifelse(Y3==1, NA, L3 + 0.1*UL4),
           A4 = ifelse(Y3==1, NA,round(as.numeric(UA4 + 0.1*L4),0)),
           Y4 = ifelse(Y3==1, 1, as.numeric((UY4+0.01*A3-0.05*L3)< 0.03)),
           A5 = ifelse(Y4==1, NA,round(as.numeric(UA5 + 0.1*L4),0)),
           C5 = ifelse(A5 == 0, 1, 0),
           Y5 = ifelse(Y4==1, 1, as.numeric((UY5+0.01*A4-0.05*L4)< 0.03)))

  #bound visits to 4 max per year
  O$A1[O$A1>2]<-3
  O$A2[O$A2>2]<-3
  O$A3[O$A3>2]<-3
  O$A4[O$A4>2]<-3

  return(O)
}

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

Cases of HCC in the Simulated Data

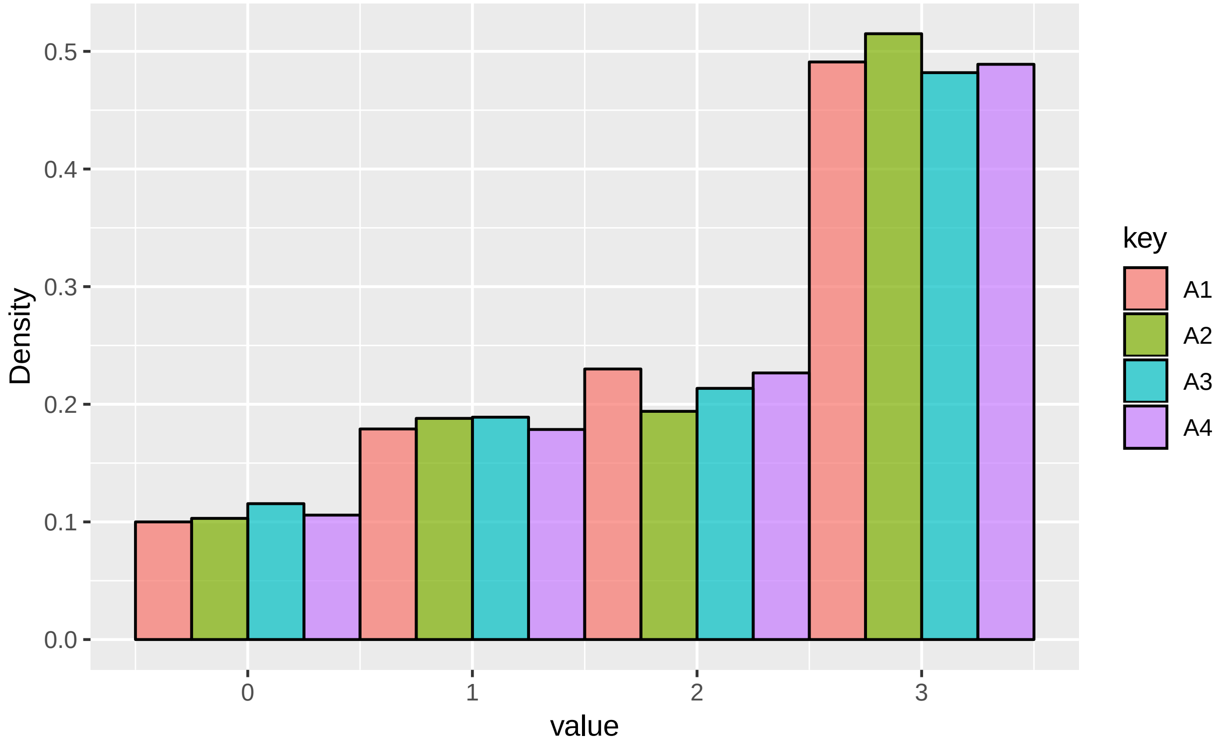
In our simulated dataset, there were 474 cases of hepatocellular carcinoma (HCC) by the end of the study interval.

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.



b. 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.

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

We simulated counterfactual outcomes with all possible combinations of 0-3 visits per year to try to determine $\Psi^F(P_{U,X})$ as the coefficient on the total number of visits in the marginal structural model $\Psi^F(P_{U,X}) =$ -0.005.

Note that some individuals have more than 3 visits per year but we felt that 3 or more indicated being well-connected to care, and this still results in 256 possible treatment combinations.

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

This means that for every additional primary care or hepatology visit (up to three annually), patients are are 0.5% less likely to develop HCC than people who have one fewer primary care or hepatology visits in the five years under study.

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

```
B = 50
estimates_data2 <- matrix(nrow = B, ncol = 3)
colnames(estimates_data2) <- c("Gcomp", "IPTW",  "TMLE")

#version with 4 levels for each A
x <- list(0:1)
```

```
abar<- expand.grid(rep(x, 8))


#set n equal to number of rows
n = as.numeric(nrow(df))

n2 <- as.numeric(nrow(abar))


#initialize sumA
sumA= rep(NA, n2)

# initialize regimes with dim n X num Anodes = 4 X num regimes
regimes = array(NA, dim=c(n,8,n2))

#fill in regimes and sumA
for(i in 1:n2){
  regimes[,,i] <- matrix(rep(as.numeric(abar[i,]),n), byrow=TRUE, nrow=n)
  sumA[i] =rowSums(abar[i,],  na.rm = TRUE)
}

#initialize and define summary measures
summary.measures = array(dim=c(n2,1,1))
dimnames(summary.measures)[[2]]=list("sumA")
summary.measures[,,1]=as.matrix(sumA)


for(b in 1:B){
df <-generate_data(1000)

#define working.msm = character formula for working MSM
working.msm = "Y ~ sumA"


#Turning As into sets of binary indicators-armadillo
#A1 first
df$A11 <- ifelse(df$A1 == 0, 0, NA)
df$A12 <- ifelse(df$A1 == 0, 0, NA)

df$A11 <- ifelse(df$A1 == 1, 1, df$A11)
df$A12 <- ifelse(df$A1 == 1, 0, df$A12)

df$A11 <- ifelse(df$A1 == 2, 0, df$A11)
df$A12 <- ifelse(df$A1 == 2, 1, df$A12)

df$A11 <- ifelse(df$A1 == 3, 1, df$A11)
df$A12 <- ifelse(df$A1 == 3, 1, df$A12)

#creating the A2 indicators
df$A21 <- ifelse(df$A2 == 0, 0, NA)
df$A22 <- ifelse(df$A2 == 0, 0, NA)

df$A21 <- ifelse(df$A2 == 1, 1, df$A21)
```

```r
df$A22 <- ifelse(df$A2 == 1, 0, df$A22)

df$A21 <- ifelse(df$A2 == 2, 0, df$A21)
df$A22 <- ifelse(df$A2 == 2, 1, df$A22)

df$A21 <- ifelse(df$A2 == 3, 1, df$A21)
df$A22 <- ifelse(df$A2 == 3, 1, df$A22)

#creating the A3 indicators
df$A31 <- ifelse(df$A3 == 0, 0, NA)
df$A32 <- ifelse(df$A3 == 0, 0, NA)

df$A31 <- ifelse(df$A3 == 1, 1, df$A31)
df$A32 <- ifelse(df$A3 == 1, 0, df$A32)

df$A31 <- ifelse(df$A3 == 2, 0, df$A31)
df$A32 <- ifelse(df$A3 == 2, 1, df$A32)

df$A31 <- ifelse(df$A3 == 3, 1, df$A31)
df$A32 <- ifelse(df$A3 == 3, 1, df$A32)

#creating the A4 indicators
df$A41 <- ifelse(df$A4 == 0, 0, NA)
df$A42 <- ifelse(df$A4 == 0, 0, NA)

df$A41 <- ifelse(df$A4 == 1, 1, df$A41)
df$A42 <- ifelse(df$A4 == 1, 0, df$A42)

df$A41 <- ifelse(df$A4 == 2, 0, df$A41)
df$A42 <- ifelse(df$A4 == 2, 1, df$A42)

df$A41 <- ifelse(df$A4 == 3, 1, df$A41)
df$A42 <- ifelse(df$A4 == 3, 1, df$A42)


df <- df %>% dplyr::select(L1, A11, A12,
                           L2, A21, A22, Y2,
                           L3, A31, A32, Y3,
                           L4, A41, A42,Y4,
                           C5,Y5)

#Rescale Ls
maxL1 <-max(df$L1)
df$L1<-ifelse(is.na(df$L1), NA, (maxL1-df$L1)/maxL1)

maxL2 <-max(df$L2)
df$L2<-ifelse(is.na(df$L2), NA,(maxL2-df$L2)/maxL2)

maxL3 <-max(df$L3,na.rm = T)
df$L3<-ifelse(is.na(df$L3), NA, (maxL3-df$L3)/maxL3)

maxL4 <-max(df$L4, na.rm = T)
df$L4<-ifelse(is.na(df$L4), NA,(maxL4-df$L4)/maxL4)
```

```
#estimate parameter using ltmleMSM function-
results2.MSM <- ltmle::ltmleMSM(df,
                     Anodes= c("A11", "A12",
                               "A21", "A22",
                               "A31", "A32",
                               "A41", "A42" ),
                     Lnodes = c("L1", "L2", "L3", "L4"),
                     Ynodes = c("Y2", "Y3", "Y4", "Y5"),
                     Cnodes = "C5",
                     working.msm = working.msm,
                     regimes = regimes,
                     summary.measures = summary.measures,
                     msm.weights = NULL,
                     survivalOutcome = TRUE)

sum.results2.MSM.tmle= summary(results2.MSM, "tmle")
sum.results2.MSM.tmle

sum.results2.MSM.iptw= summary(results2.MSM, "iptw")
sum.results2.MSM.iptw

results2.MSM.g <- ltmle::ltmleMSM(df,
                     Anodes= c("A11", "A12",
                               "A21", "A22",
                               "A31", "A32",
                               "A41", "A42"),
                     Lnodes = c("L1", "L2", "L3", "L4"),
                     Ynodes = c("Y2", "Y3", "Y4", "Y5"),
                     Cnodes = "C5",
                     working.msm = working.msm,
                     regimes = regimes,
                     summary.measures = summary.measures,
                     msm.weights = NULL,
                     survivalOutcome = TRUE,
                     gcomp = TRUE)

sum.results2.MSM.g = summary(results2.MSM.g, "gcomp")
sum.results2.MSM.g

estimates_data2[b,] <- c(sum.results2.MSM.g$cmat[2,1],
                         sum.results2.MSM.iptw$cmat[2,1],
                         sum.results2.MSM.tmle$cmat[2,1])
print(B)

}


Bias.g <- mean(estimates_data2[,1]) - PsiF
Variance.g <- var(estimates_data2[,1])
MSE.g <- mean((estimates_data2[,1]- PsiF)^2)

Bias.iptw <- mean(estimates_data2[,2]) - PsiF
Variance.iptw <- var(estimates_data2[,2])
```

```r
MSE.iptw <- mean((estimates_data2[,2]- PsiF)^2)

Bias.tmle <- mean(estimates_data2[,3]) - PsiF
Variance.tmle <- var(estimates_data2[,3])
MSE.tmle <- mean((estimates_data2[,3]- PsiF)^2)
```

Unfortunately, this simulation ran for >60 hours and ultimately crashed Stephanie's laptop (without saving :/). We'd love some help on how to optomize this. Notably we keep getting an error message about rescaling the L variables (even after we thought we rescaled them), so that might be one thing to look at. We realize that one option is to just binarize our exposure (we do that below to try the analysis, though we haven't done the binary estimator performance yet) but we were really hoping to be able to capture more visit data.

We did save the results of a single iteration of the simulated estimator performance check and got the following from that single iteration:

d. Data example

```r
#Turning observed data into something that we can analyze
O <- as.data.frame(matrix(nrow=nrow(HCVData), NA))
O$Y1 <- ifelse(HCVData$Cancer.Diagnosis.Date>'2015-12-31' |
                 is.na(HCVData$Cancer.Diagnosis.Date), 0, 1)
O$Y2 <- ifelse(HCVData$Cancer.Diagnosis.Date>'2016-12-31' |
                 is.na(HCVData$Cancer.Diagnosis.Date), 0, 1)
O$Y3 <- ifelse(HCVData$Cancer.Diagnosis.Date>'2017-12-31' |
                 is.na(HCVData$Cancer.Diagnosis.Date), 0, 1)
O$Y4 <- ifelse(HCVData$Cancer.Diagnosis.Date>'2018-12-31' |
                 is.na(HCVData$Cancer.Diagnosis.Date), 0, 1)
O$Y5 <- ifelse(HCVData$Cancer.Diagnosis.Date>'2019-12-31' |
                 is.na(HCVData$Cancer.Diagnosis.Date), 0, 1)

O$L1 <- HCVData$FIB4_by2015
O$L2 <- HCVData$FIB4_by2016
O$L3 <- HCVData$FIB4_by2017
O$L4 <- HCVData$FIB4_by2018

O$L1lm <- HCVData$YearsSinceFib4_2015
O$L2lm <- HCVData$YearsSinceFib4_2016
O$L3lm <- HCVData$YearsSinceFib4_2017
O$L4lm <- HCVData$YearsSinceFib4_2018

O$A1 <- HCVData$VISIT_COUNT_2015
O$A2 <- HCVData$VISIT_COUNT_2016
O$A3 <- HCVData$VISIT_COUNT_2017
O$A4 <- HCVData$VISIT_COUNT_2018
O$A5 <- HCVData$VISIT_COUNT_2019

#truncate visits to 3
O$A1[O$A1>2] <-3
O$A2[O$A2>2] <-3
O$A3[O$A3>2] <-3
O$A4[O$A4>2] <-3

O$C5 <- ifelse(O$A5 == 0, 1, 0)
```

```
O <- O %>% dplyr::select(-V1)

#ESTIMATE
df <- O

#version with 4 levels for each A
x <- list(0:1)
abar<- expand.grid(rep(x, 8))

#set n equal to number of rows
n <- as.numeric(nrow(df))

n2 <- as.numeric(nrow(abar))

# #COMMENTING OUT SO DOCUMENT WILL KNIT WITHOUT RUNNING AGAIN
# #initialize sumA
# sumA= rep(NA, n2)
#
# # initialize regimes with dim n X num Anodes = 4 X num regimes
# regimes = array(NA, dim=c(n,8,n2))
#
# #fill in regimes and sumA
# for(i in 1:n2){
#   regimes[,,i] <- matrix(rep(as.numeric(abar[i,]),n), byrow=TRUE, nrow=n)
#   sumA[i] =rowSums(abar[i,],  na.rm = TRUE)
# }
#
# #initialize and define summary measures
# summary.measures = array(dim=c(n2,1,1))
# dimnames(summary.measures)[[2]]=list("sumA")
# summary.measures[,,1]=as.matrix(sumA)
#
#
# #define working.msm = character formula for working MSM
# working.msm = "Y ~ sumA"
#
# #Turning As into sets of binary indicators-armadillo
# #A1 first
# df$A11 <- ifelse(df$A1 == 0, 0, NA)
# df$A12 <- ifelse(df$A1 == 0, 0, NA)
#
# df$A11 <- ifelse(df$A1 == 1, 1, df$A11)
# df$A12 <- ifelse(df$A1 == 1, 0, df$A12)
#
# df$A11 <- ifelse(df$A1 == 2, 0, df$A11)
# df$A12 <- ifelse(df$A1 == 2, 1, df$A12)
#
# df$A11 <- ifelse(df$A1 == 3, 1, df$A11)
# df$A12 <- ifelse(df$A1 == 3, 1, df$A12)
#
# #creating the A2 indicators
# df$A21 <- ifelse(df$A2 == 0, 0, NA)
# df$A22 <- ifelse(df$A2 == 0, 0, NA)
```

```
#
# df$A21 <- ifelse(df$A2 == 1, 1, df$A21)
# df$A22 <- ifelse(df$A2 == 1, 0, df$A22)
#
# df$A21 <- ifelse(df$A2 == 2, 0, df$A21)
# df$A22 <- ifelse(df$A2 == 2, 1, df$A22)
#
# df$A21 <- ifelse(df$A2 == 3, 1, df$A21)
# df$A22 <- ifelse(df$A2 == 3, 1, df$A22)
#
# #creating the A3 indicators
# df$A31 <- ifelse(df$A3 == 0, 0, NA)
# df$A32 <- ifelse(df$A3 == 0, 0, NA)
#
# df$A31 <- ifelse(df$A3 == 1, 1, df$A31)
# df$A32 <- ifelse(df$A3 == 1, 0, df$A32)
#
# df$A31 <- ifelse(df$A3 == 2, 0, df$A31)
# df$A32 <- ifelse(df$A3 == 2, 1, df$A32)
#
# df$A31 <- ifelse(df$A3 == 3, 1, df$A31)
# df$A32 <- ifelse(df$A3 == 3, 1, df$A32)
#
# #creating the A4 indicators
# df$A41 <- ifelse(df$A4 == 0, 0, NA)
# df$A42 <- ifelse(df$A4 == 0, 0, NA)
#
# df$A41 <- ifelse(df$A4 == 1, 1, df$A41)
# df$A42 <- ifelse(df$A4 == 1, 0, df$A42)
#
# df$A41 <- ifelse(df$A4 == 2, 0, df$A41)
# df$A42 <- ifelse(df$A4 == 2, 1, df$A42)
#
# df$A41 <- ifelse(df$A4 == 3, 1, df$A41)
# df$A42 <- ifelse(df$A4 == 3, 1, df$A42)
#
#
# df <- df %>% dplyr::select(L1, A11, A12,
#                             L2, A21, A22, Y2,
#                             L3, A31, A32, Y3,
#                             L4, A41, A42,Y4,
#                             C5,Y5)
#
# #Rescale Ls
# maxL1 <-max(df$L1)
# df$L1<-ifelse(is.na(df$L1), NA, (maxL1-df$L1)/maxL1)
#
# maxL2 <-max(df$L2)
# df$L2<-ifelse(is.na(df$L2), NA,(maxL2-df$L2)/maxL2)
#
# maxL3 <-max(df$L3,na.rm = T)
# df$L3<-ifelse(is.na(df$L3), NA, (maxL3-df$L3)/maxL3)
#
```

```
# maxL4 <-max(df$L4, na.rm = T)
# df$L4<-ifelse(is.na(df$L4), NA,(maxL4-df$L4)/maxL4)
#
# #estimate parameter using ltmleMSM function-
# results2.MSM <- ltmle::ltmleMSM(df,
#                       Anodes= c("A11", "A12",
#                                 "A21", "A22",
#                                 "A31", "A32",
#                                 "A41", "A42" ),
#                       Lnodes = c("L1", "L2", "L3", "L4"),
#                       Ynodes = c("Y2", "Y3", "Y4", "Y5"),
#                       Cnodes = "C5",
#                       working.msm = working.msm,
#                       regimes = regimes,
#                       summary.measures = summary.measures,
#                       msm.weights = NULL,
#                       survivalOutcome = TRUE)
#
# sum.results2.MSM.tmle= summary(results2.MSM, "tmle")
# sum.results2.MSM.tmle
#
# sum.results2.MSM.iptw= summary(results2.MSM, "iptw")
# sum.results2.MSM.iptw
#
# results2.MSM.g <- ltmle::ltmleMSM(df,
#                       Anodes= c("A11", "A12",
#                                 "A21", "A22",
#                                 "A31", "A32",
#                                 "A41", "A42"),
#                       Lnodes = c("L1", "L2", "L3", "L4"),
#                       Ynodes = c("Y2", "Y3", "Y4", "Y5"),
#                       Cnodes = "C5",
#                       working.msm = working.msm,
#                       regimes = regimes,
#                       summary.measures = summary.measures,
#                       msm.weights = NULL,
#                       survivalOutcome = TRUE,
#                       gcomp = TRUE)
#
# sum.results2.MSM.g = summary(results2.MSM.g, "gcomp")
# sum.results2.MSM.g
#
# estimates_0 <- c(sum.results2.MSM.g$cmat[2,1],
#                           sum.results2.MSM.iptw$cmat[2,1],
#                           sum.results2.MSM.tmle$cmat[2,1])
```

```
#binary version of the analysis (yes/no visits each year)
x <- list(0:1)

abar<- expand.grid(rep(x, 4))

df <- 0

#set n equal to number of rows
```

```
n <- as.numeric(nrow(df))

n2 <- as.numeric(nrow(abar))

#truncate visits to 1
df$A1[df$A1>0] <-1
df$A2[df$A2>0] <-1
df$A3[df$A3>0] <-1
df$A4[df$A4>0] <-1

#initialize sumA
sumA= rep(NA, n2)

# initialize regimes with dim n X num Anodes = 4 X num regimes
regimes = array(NA, dim=c(n,4,n2))

#fill in regimes and sumA
for(i in 1:n2){
  regimes[,,i] <- matrix(rep(as.numeric(abar[i,]),n), byrow=TRUE, nrow=n)
  sumA[i] =rowSums(abar[i,],  na.rm = TRUE)
}

#initialize and define summary measures
summary.measures = array(dim=c(n2,1,1))
dimnames(summary.measures)[[2]]=list("sumA")
summary.measures[,,1]=as.matrix(sumA)

# #define working.msm = character formula for working MSM
working.msm = "Y ~ sumA"

df <- df %>% dplyr::select(L1, A1,
                           L2, A2, Y2,
                           L3, A3, Y3,
                           L4, A4, Y4,
                           C5, Y5)

#Rescale Ls
maxL1 <-max(df$L1)
df$L1<-ifelse(is.na(df$L1), NA, (maxL1-df$L1)/maxL1)

maxL2 <-max(df$L2)
df$L2<-ifelse(is.na(df$L2), NA,(maxL2-df$L2)/maxL2)

maxL3 <-max(df$L3,na.rm = T)
df$L3<-ifelse(is.na(df$L3), NA, (maxL3-df$L3)/maxL3)

maxL4 <-max(df$L4, na.rm = T)
df$L4<-ifelse(is.na(df$L4), NA,(maxL4-df$L4)/maxL4)

#estimate parameter using ltmleMSM function-
resultsbin.MSM <- ltmle::ltmleMSM(df,
                     Anodes = c("A1", "A2", "A3", "A4"),
                     Lnodes = c("L1", "L2", "L3", "L4"),
                     Ynodes = c("Y2", "Y3", "Y4", "Y5"),
                     Cnodes = "C5",
                     working.msm = working.msm,
                     regimes = regimes,
                     summary.measures = summary.measures,
                     msm.weights = NULL,
                     survivalOutcome = TRUE)

sum.resultsbin.MSM.tmle= summary(resultsbin.MSM, "tmle")
sum.resultsbin.MSM.tmle
```

```
## Estimator:  tmle
##             Estimate Std. Error  CI 2.5% CI 97.5%  p-value
## (Intercept) -1.06729    0.11711 -1.29681   -0.838  < 2e-16 ***
## sumA        -0.20939    0.04009 -0.28796   -0.131 1.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sum.resultsbin.MSM.iptw= summary(resultsbin.MSM, "iptw")
sum.resultsbin.MSM.iptw
```

```
## Estimator:  iptw
##             Estimate Std. Error  CI 2.5% CI 97.5% p-value
## (Intercept) -1.36931    0.15669 -1.67642   -1.062  <2e-16 ***
## sumA        -0.12204    0.05468 -0.22921   -0.015  0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
resultsbin.MSM.g <- ltmle::ltmleMSM(df,
                     Anodes = c("A1", "A2", "A3", "A4"),
                     Lnodes = c("L1", "L2", "L3", "L4"),
                     Ynodes = c("Y2", "Y3", "Y4", "Y5"),
                     Cnodes = "C5",
                     working.msm = working.msm,
                     regimes = regimes,
                     summary.measures = summary.measures,
                     msm.weights = NULL,
                     survivalOutcome = TRUE,
                     gcomp = TRUE)

sum.resultsbin.MSM.g = summary(resultsbin.MSM.g, "gcomp")
sum.resultsbin.MSM.g
```

```
## Estimator:  gcomp
## Warning: inference for gcomp is not accurate! It is based on TMLE influence curves.
##             Estimate Std. Error  CI 2.5% CI 97.5%  p-value
## (Intercept) -1.16795    0.12306 -1.40913   -0.927  < 2e-16 ***
## sumA        -0.19929    0.04186 -0.28133   -0.117 1.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
estimates_Obin <- c(sum.resultsbin.MSM.g$cmat[2,1],
                         sum.resultsbin.MSM.iptw$cmat[2,1],
                         sum.resultsbin.MSM.tmle$cmat[2,1])

save(estimates_Obin, file = "estimates_Obin112319.Rdata")
```

Ultimately, implementing *ltmle* on our observed data produced the following estimates for the β on $\bar{A}(t) = \bar{a}(t)$ where $a \in \{0, 1, 2, 3+\}$ hepatology or primary care visits in each timepoint (i.e. per year):

G-Comp	IPTW	TMLE
-0.099	0.17	-0.134

If we use the TMLE estimate (an arbitrary choice at this point given that we could not complete our simulation and compare performance of the various estimators), this means that for every additional primary care or hepatology visit (up to three annually), patients are are 13.4% less likely to develop HCC than people who have one fewer primary care or hepatology visits in the five years under study, adjusting for liver cirrhosis.

Implementing *ltmle* on our observed data produced the following estimates for the β on $\bar{A}(t) = \bar{a}(t)$ where $a \in \{0, 1+\}$ hepatology or primary care visits in each timepoint (i.e. per year) - a binary outcome at each A(t):

G-Comp	IPTW	TMLE
-0.199	-0.122	-0.209

If we use the TMLE estimate, this means that over the five years under study, patients are 20.9% less likely to develop HCC each year that they had at least one primary care or hepatology visit, compared to patients who had no primary care of hepatology visits that year, adjusting for liver cirrhosis.

e. "None of the options are good", and other life lessons from Causal

Despite our efforts to simply our analysis to limit the number of possible treatment regimes (i.e. examining the effect of 0, 1, 2, and 3 or more visits, instead of the full distribution of visit counts), this analysis is quite competationally demanding; *ltmle* is extremely slow, with 1 iteration of the simulation taking more than an hour to complete and 50 iterations running for >60 hours and crashing Steph's machine. We will continue to proceed and plan time for R to run analyses as needed, but any suggestions for how to streamline would be welcome!!

f. Next steps

We *just* received an updated version of the dataset with diagnosis dates, while our simulation was already running (even though that was ultimately fruitless anyway). Besides trying to streamline our simulation and estimation code, we also plan to tune our simulation to be closer to our real data given that we know know the HCC diagnosis dates. We will then check estimator performance with our simulation, which will help us better interpret the findings from the implementation of those estimators on our observed data.

5. References

Axley P, Ahmed Z, Ravi S, Singal AK. Hepatitis C Virus and Hepatocellular Carcinoma: A Narrative Review. J Clin Transl Hepatol. 2018 Mar 28; 6(1): 79-84.

Bosetti C, Turati F, La Vecchia C. Hepatocellular carcinoma epidemiology.Best Pract Res Clin Gastroenterol. 2014 Oct; 28(5):753-70.

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

Burstow NJ, Mohamed Z, Gomaa AI, et al. Hepatitis C treatment: where are we now? Int J Gen Med 2017; 10:39-52.

Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med. 2003 Apr 24; 348(17):1625-38.

Chuang SC, Lee YC, Hashibe M, Dai M, Zheng T, Boffetta P. Interaction between cigarette smoking and hepatitis B and C virus infection on the risk of liver cancer: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2010 May; 19(5):1261-8.

Donato F, Tagger A, Gelatti U, Parrinello G, Boffetta P, Albertini A, Decarli A, Trevisi P, Ribero ML, Martelli C, Porru S, Nardi G. Alcohol and hepatocellular carcinoma: the effect of lifetime intake and hepatitis virus infections in men and women. Am J Epidemiol. 2002 Feb 15; 155(4):323-31.

El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology. 2012 May; 142(6):1264-1273.e1.

European Association for Study of Liver. EASL Recommendations on Treatment of Hepatitis C 2015. J Hepatol. 2015 Jul; 63(1):199-236.

Huang TS, Lin CL, Lu MJ, Yeh CT, Liang KH, Sun CC, Shyu YC, Chien RN. Diabetes, hepatocellular carcinoma, and mortality in hepatitis C-infected patients: A population-based cohort study. J Gastroenterol Hepatol. 2017 Jul; 32(7):1355-1362.

Khan MQ, Anand V, Hessefort N, Hassan A, Ahsan A, Sonnenberg A, Fimmel CJ. Utility of Electronic Medical record-based Fibrosis Scores in Predicting Advanced Cirrhosis in Patients with Hepatitic C Virus Infection. J Transl Int Med. 2017 Mar; 5(1): 43-48.

Omland LH, Krarup H, Jepsen P, Georgsen J, Harritsh?j LH, Riisom K, Jacobsen SE, Schouenborg P, Christensen PB, S?rensen HT, Obel N, DANVIR Cohort Study. Mortality in patients with chronic and cleared hepatitis C viral infection: a nationwide cohort study. J Hepatol. 2010 Jul; 53(1):36-42.

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