STAT 309 Final Project Write-Up - Karolinska Institute Cardia Cancer Data

Introduction

The purpose of this project is to assess whether being treated at a hospital that sees a high volume of patients with cardia cancer leads to different survival rates than being treated at a low-volume hospital. A problem that arises when we try to answer this question is that the hospital where patients were initially diagnosed with cardia cancer is not always the same as the hospital where they receive most of their medical treatment. To address this, I am using an instrumental variables framework. I do not think that the hospital patients are treated in is unconfounded, because there are many factors that go into a person's decision to be treated at a high vs low volume hospital. However, I think that the hospital people are diagnosed in is much more likely to be unconfounded, since people are probably going to the hospital they usually go to and is closest to them, and so we have the appropriate covariates to assume unconfoundedness. Therefore, I designed my study based on the diagnosing hospital, and used it to learn about the treating hospital.

Assumptions

There are several assumptions that need to be addressed before going further with the project. I believe SUTVA I holds because we are only looking at high-volume or non high-volume hospitals, so there are no different forms of treatment that would lead to different potential outcomes. However, I am worried about SUTVA II holding. The hospital a patient is being treated at could affect the outcome of another patient being treated there if there are limited resources. This could be an issue, but I continued with my analysis as if SUTVA II does hold. There are also assumptions specific to the instrumental variables method-of-moments analysis. The monotonicity assumption is that there are no defiers. In this context, it means there are no

patients who if diagnosed in a high-volume hospital will get treated at a low-volume hospital, but if diagnosed at a low-volume hospital will get treated in a high-volume hospital. I am worried about this assumption holding, because it seems possible that there are people who would always want to be treated at a hospital different from the one they were diagnosed in. For example, they might associate the type of hospital they were diagnosed in with bad memories because they found out they had cancer there, so they would choose a different type of hospital to be treated in. Other assumptions for the method-of-moments analysis are the exclusion restrictions. Always-takers are people who will get treated at a high-volume hospital no matter where they are diagnosed. The exclusion restriction for always-takers states that the causal effect of diagnosing an always-taker at a high vs low volume hospital is zero. Never-takers are people who will get treated at a low-volume hospital no matter where they are diagnosed. The exclusion restriction for never-takers is that the causal effect of diagnosing a never-taker at a high vs low volume hospital is zero. I think it is likely that the exclusion restriction assumptions hold in this context.

Method

Before beginning to create subclasses, I created several new variables. I used the DateOfDiagnosis variable to create a variable called MonthOfDiagnosis. I later removed MonthOfDiagnosis when checking for covariate balance, because it labeled the dates numerically 1-12, which makes January and December appear as very different even though they should not be. I also used MonthOfDiagnosis to make another variable called GoodWeather, which is a binary variable that has a 1 if the patient was diagnosed between the months of May and September, and 0 otherwise. I created this variable because Swedish weather is very different during different times of the year, and I thought it might have an impact on the hospital that

patients chose to be treated in. For example, during bad weather there is more likely to be a lot of snow, which could prevent someone from going to get treated at a hospital that is farther away from them, or during good weather a patient's family might be on vacation and would not be able to help the patient get to the hospital they prefer. I decided May-September as the months with good weather, because on average those are the months that are warmest and most sunny in Sweden (Climate and Average Weather Year Round in Stockholm, n.d.). I also created a variable called OldRuralMale which is a binary indicator for a patient who is from a rural area, male, and over the age of 65. I used 65 as the cutoff because that is the average retirement age in Sweden (Tuck, 2021).

I tried several different subclass designs and checked their covariate balance to choose the one I thought was best. First, I tried a logistic regression model using all the covariates in the dataset to get propensity scores. I did not include any specific interactions in my model because I thought having the OldRuralMale variable would be enough. I discarded the 16 units with a propensity score below 0.17 because they were all in control and had no matches in active. All of these units are old, rural, males, and diagnosed during good weather. Removing them from my analysis means that I will not be able to generalize my conclusions to this population. I then created four subclasses based on the units' propensity scores, and checked the covariate balance in this design by getting the means of the covariates in each subclass and weighing them by the number of units in each subclass. Next, I tried a different logistic regression model on the covariates AgeAtDiagnosis, FromRuralArea, Male, and DateOfDiagnosis, because I thought that the balance on these covariates were especially important. I discarded the 3 units that had a propensity score below 0.20, who were all old rural males and had no equivalents in the active group. Like before, I created subclasses based on propensity scores and checked covariate

balance. Finally, I tried a tree model on all the covariates. The tree only had one split, on the variable OldRuralMale, and I used the resulting propensity scores to create subclasses. Below is a table showing the original covariate balance compared with the covariate balance I got for each of the three subclass designs (the design with the best balance for a covariate is highlighted).

	Original data (No subclasses)		Subclass design 1 (logistic reg.)		Subclass design 2 (logistic reg.)		Subclass design 3 (tree)	
	$\overline{X_T}$	$\overline{X_C}$	$\overline{X_T}$	$\overline{X_C}$	$\overline{X_T}$	$\overline{X_C}$	$\overline{X_T}$	$\overline{X_C}$
Age At Diagnosis	64.66	68.43	65.71	66.43	66.54	66.37	64.33	67.66
From Rural Area	0.37	0.67	0.46	0.49	0.51	0.48	0.52	0.52
Male	0.68	0.81	0.69	0.72	0.74	0.77	0.66	0.80
Date of Diagnosis	8064.33	8033.34	8072.91	8031.79	8090.33	8047.85	8079.09	8022.97
Good Weather	0.32	0.42	0.31	0.26	0.25	0.38	0.29	0.40
Old Rural Male	0.09	0.39	0.15	0.15	0.24	0.22	0.12	0.30

The first logistic regression model had the best balance on the most covariates, and was balanced well on the other covariates as well, so I chose it as my study design. Using diagnosing hospital as an instrument, I then estimated the average causal effect of high vs low volume treating hospital on survival using the method-of-moments analysis. First, I created a binary outcome variable for survival, where surviving longer than one year after diagnosis is a 1 and surviving only one year after diagnosis is a 0. I decided to make this the outcome variable, because cancer is very deadly so I thought it would be valuable to a patient to know whether they can be expected to survive for more than one year or not. I calculated the estimated

intention-to-treat effect ($\widehat{ITT}$) for each subclass by subtracting the mean survival of patients diagnosed in a low-volume hospital from the mean survival of patients diagnosed in a high-volume diagnosing hospital in each subclass. To calculate the estimated proportion of compliers ($\widehat{p_C}$), I first calculated the estimated proportion of always-takers ($\widehat{p_A}$) and proportion of compliers or always-takers ($\widehat{p_{C\ or\ A}}$). Patients who were diagnosed in low-volume and treated in high-volume hospitals are either always-takers or defiers. With the monotonicity assumption I am assuming there are no defiers, so the proportion of people who fit this description are the $\widehat{p_A}$. Patients who were both diagnosed and treated in high-volume hospitals are either compliers or always-takers, so if I calculate the proportion of such patients I get $\widehat{p_{C\ or\ A}}$. To get $\widehat{p_C}$ I calculate $\widehat{p_{C\ or\ A}} - \widehat{p_A}$ in each subclass. To get the estimated intention-to-treat effect for compliers ($\widehat{ITT_C}$), I calculate $\widehat{ITT}/\widehat{p_C}$ in each subclass, then I multiply the results by the weights I calculated earlier.

Results

My resulting total $\widehat{ITT_C}$ is 0.11. This is the causal effect of being treated at a high vs low volume hospital on surviving more than one year after diagnosis for the type of people who are compliers. This indicates that compliers who get treated at a high volume hospital might be more likely to survive longer than one year. To get the interval for the estimate, I calculated the variance for $\widehat{ITT_C}$ in each subclass and multiplied the results by the weights squared. Thus, I calculated that the 95% confidence interval for the $\widehat{ITT_C}$ is from -0.27 to 0.47.

Discussion

Stage of cancer was removed from this data set because it might not be considered a pre-treatment covariate. This is because the staging of cancer can be affected by where a patient

is diagnosed, for example a high volume hospital might have more experience with accurately assessing the stage of cancer. Before beginning the study, the observed covariates that I thought were most likely to predict the treatment decisions was FromRuralArea, because a person's location is going to affect their distance to a high or low volume hospital, which would affect their decision. I also thought that the AgeOfDiagnosis variable would be important in predicting both treatment decisions and outcomes, because an older person might be less inclined to travel a long distance to go to a specific hospital they want, and a patient's age might affect their likelihood of survival. I thought that the combination of an old, rural, male patient could be important, so I created a variable for it to include in my design. I also believed that the time of year a patient was diagnosed in could affect their decision so I included a variable for that as well. The fact that this data was collected from Sweden affects which covariates I believe are important. For example, if the data was collected in the U.S., covariates relating to income, socioeconomic status, and healthcare coverage would be very important due to the way our healthcare system functions. However, Sweden has a universal healthcare system where all residents can easily afford medical services, so those covariates are not as relevant (Tikkanen et al., 2020). Additionally, if the data had been collected from a country that had similar weather all year round, I would think it's less likely that the time of year would have an effect on people's treatment decisions. But because Sweden has very different weather throughout the year, I decided to create a variable for good weather. I don't think these covariates are enough to assume unconfoundedness in treatment hospital decisions, because there are many factors that would go into making that decision. However, I do think these covariates are enough to assume unconfoundedness for diagnosing hospitals, as the patients don't know they are going to get diagnosed with cancer when they go to the hospital. Unconfoundedness is important because it

determines which questions we can try to answer with the data and what kind of design we need to use. We cannot design an observational study with the treatment hospital as the treatment intervention because we do not have enough covariates to assume unconfoundedness. So instead we use the instrumental variables framework.

Citations

Climate and Average Weather Year Round in Stockholm. Stockholm Climate, Weather By Month, Average Temperature (Sweden). (n.d.). https://weatherspark.com/y/84156/Average-Weather-in-Stockholm-Sweden-Year-Round

Tikkanen, R., Osborn, R., Mossialos, E., Djordjevic, A., & Wharton, G. A. (2020, June 5). *International Healthcare System Profiles - Sweden*. The Commonwealth Fund. https://www.commonwealthfund.org/international-health-policy-center/countries/sweden

Tuck, N. (2021, April 29). *Average retirement age for Swedes rises to 65 – Swedish Pensions Agency*. European Pensions. https://www.europeanpensions.net/ep/Average-retirement-age-for-Swedes-rises-to-65-Swedish-Pensions-Agency.php