Matthew Guest Grant Final Research Report

Klaas van Kempen

Donna Gilleskie (Mentor)

September 20, 2019

Research Question:

The aim of my summer research project is to understand the roles of observable patient-level medical care-related behaviors on the evolution of health markers measuring kidney function among adolescents with chronic kidney disease (CKD). The health markers of interest include estimated glomerular filtration rate (eGFR), blood pressure, and proteinuria and albuminuria. The medical care-related behaviors include adherence to recommended clinic appointments, medication adherence, and emergency room visits stemming from health shocks.

This research question is part of a larger study, named the LIFE COURSE Study (Longitudinal Indicators For Evaluating Clinical Outcomes with Underlying Renal disease in a Sample of Emerging adults) with Dr. Maria Ferris as lead investigator and Dr. Donna Gilleskie as a co-investigator. I will have access to data from the LIFE COURSE Study and Dr. Gilleskie will serve as my advisor.

Data Sets:

In order to address our research question we had access to 10 types of data obtained from electronic medical records stored in two different UNC EMR systems: Legacy and Epic. The data include: demographics, appointments, encounters, charges, diagnoses, labs, vitals, hospital procedures, and home addresses. Each of the data sets contains longitudinal information on 636 patients, 312 of which were identified as being CKD patients. The vast array of available data allows for information on demographics and socioeconomic status, as well as data on adherence to appointments, health shocks, medications taken, and importantly lab values and vitals.

Data Cleaning:

In order to address our research question we first had to manipulate the multi-level data to get it in a usable state. With 10 files to work with, and over 1 million unique observations in some of them, the bulk of my summer was spent cleaning the data. This task involved evaluating responses, fixing errors, and aggregating responses to generate new variables when necessary.

Merging the Data:

Once each of the 10 data files were appropriately cleaned, we focused on how best to merge the various levels of data into a longitudinal picture of each patient's life course of events. Working closely with Dr. Gilleskie, we decided to focus our preliminary analysis on how creatinine values at a particular date/age of a patient might be explained by its previous level, time since last creatinine measure, health shocks (emergency room visits), total encounters, and various demographic variables: age, gender, ethnicity. We arranged the data to identify a creatinine value at lab date ‘t’ as the unit of analysis and gathered health related information from the last time the lab was taken and all the days in between so that for each patient, each unique date had all of the measurements and readings taken that day on the same row.

Data Analysis:

After aligning lab values by date and patient and the associated explanatory variables, we normed continuous variables and identified a reference group for polychotomous variables to aid in interpretation of coefficients in simple OLS regressions that measure how variations in explanatory variables explain variation in creatinine values. For example, our mean time between readings was 60 days and therefore we produced a normed time since last reading variable that was your time since last reading minus 60 days.

We started our analysis by running an OLS regression in STATA with creatinine values being the dependent variable and using the following variables that capture health effects between periods (a period is defined as a time period between two creatinine readings):

- Normed time since last creatinine reading
- Number of Hospital stays in last period
- Total # of Encounters in previous period minus hospital stays
- Normed Age (12 years)
- Indicator Variables for Age, Gender, Race, and Primary Language spoken at home

We found that for a reference individual who is white, female, speaks english, and is 12 years old we would expect, on average, a creatinine level of 2.778 (regression constant) :

For each additional day between creatinine readings over 60 days we would expect to see a 0.002177 drop in creatinine level today and this effect was statistically significant. Each additional encounter in the previous period suggested a 0.06915 drop in creatinine which was statistically significant. Any reported ethnicity that was not white suggested higher creatinine level and all were statistically significant. This regression did not include previous creatinine value as an explanatory variable which may explain the statistical significance for many of the variables in this regression.

Once we control for the previous creatinine level as an explanatory variable we estimate the following relationships:

We find that, each additional day past 60 days between lab readings would suggest a 0.0013 increase in creatinine. For each year of age over 12 years old we would expect to see a 0.109 increase in creatinine. Only “Other Race” and “Hispanic” individuals show a statistically significant difference in creatinine from Caucasians, and both show an expected increase in creatinine.

While this initial analysis yield some insight into what might influence creatinine levels, we know that creatinine is not normally distributed in the population even among healthy patients, rather it is skewed right, even more so among CKD patients. A standard OLS regression is not the best choice for a regression.To better analyze our data we will use a glm regression of our dependent and explanatory variables. We anticipate that marginal effects will be less biased and more efficient.

Future Analysis:

As mentioned earlier we will use alternative regression models to better analyze our data. We will also work to add in more behavioral aspects into our regression. We intend to include adherence to appointments, socioeconomic status, how often an individual relocates, or how often an individual switches insurance. How these behavioral aspects impact creatinine levels could be very useful in explaining the progression of Chronic Kidney Disease. We will also examine blood pressure and other relevant lab values such as protein and albumin.

Currently we have been using creatinine as the indicator for kidney function, however a better indicator is GFR which is a function of an individual's creatinine and height. We will derive a GFR value for our analysis so that we can better understand the progression of CKD.

Summer Experience:

Having the opportunity to work with Dr. Gilleskie on this project has been incredible. While I had other research jobs prior to working on this project, this was my first experience with data cleaning and analysis. Working with these data allowed me to become proficient in using STATA to clean and analyze data, which will be a useful skill to have for future research opportunities. I am also fortunate to have the opportunity to continue to work on the LIFECOURSE study with Dr. Gilleskie and Dr. Ferris (PI), which will give me further experience with data analysis and the possibility for future publication.

Receiving this summer research grant was also key for my development as a student. Having just finished my freshman year when I started to work on this project, my view of the economics field was very limited to what I learned in my introductory courses. Working with Dr. Gilleskie has shown me what it truly means to do economic research and I have found it incredibly interesting to see how it can be applied to different fields, including medicine. Through this opportunity I have discovered that a graduate degree in economics is definitely something that I want to pursue.