**MU Data Science Capstone – Spring 2018**

**Author(s):** David Hedrick

**Team-member specific contributions:**

I am working alone on this project, and therefore have made 100% of the contribution. I do not anticipate this to change for the duration of this project.

**Progress to Date:**

**3/7/2018:** I am working with data that I use every day in a work environment, which has afforded me the unique advantage of knowing exactly what I have access to, and the scope and limitations of that data. Most of the time spent to date has been spent on data carpentry. This was expected, as health care utilization data is quite complex and needs to be condensed into usable format for analysis. I am using a combination of Medicaid claims and eligibility data, as well as clinical documentation in the Nationwide Children’s electronic medical record to derive each member-level attribute. My final dataset is comprised of one line per Medicaid Managed Care who met certain continuous eligibility criteria during calendar year 2016. The data are now in a flat file dataset manufactured from the data sources previously mentioned.

Once the data carpentry step was completed, I moved on to conducting the analysis and began the storytelling process. In addition to laying out a background for the project in a deliverable format, I have completed the initial propensity scoring and propensity matching of the treated and non-treated populations, and I have evaluated the “closeness” of the match.

**3/28/2018:**

In addition to including all the code and output within the markdown document, I implemented some JavaScript and CSS code into the background of the html document to allow the user to toggle visibility of the R code, output and plots. I also implemented an R package that allowed me to embed an html widget to make the pre-match vs. post-match density plot section interactive. A dropdown menu is available to select either the pre-match or post-match density plot, and each plot that appears was developed using the plotly library, which allows for further user interaction.

After completing the propensity match and visually inspecting its accuracy, I tested the significance of the differences between the enrolled and non-enrolled populations for each of the covariates. Ideally, there should be no significant difference between any of the covariates. After significance testing, I moved onto evaluation of the results between the enrolled and non-enrolled populations. I also tested the significance of the differences between the two populations in the measurement data. Based on the results, I concluded that members in the care coordination do not have fewer avoidable emergency visits in the post-enrollment period.

**Issues Encountered:**

**3/7/2018:** My organization is very protective of the data, partly due to HIPAA compliance standards, but also due to other legal considerations around the use of the claims and eligibility data. Since we are capitated by the five Ohio Medicaid Managed Care Plans themselves rather than being a health plan on our own, we do not own the claims and eligibility data. Instead, the claims and eligibility data are provided to us as decision support and quality improvement resources. Consequently, I was required to submit an amendment to a data use agreement I submitted in the summer of 2017. While this caused some delay in getting the project started, I feel I have made good progress to-date and do not anticipate further issues.

**3/28/2018:** One challenge I encountered was the inability to make a good match using as many covariates as I originally proposed. I felt some covariates, such as flags for history of abuse and history of non-traditional living situations were included for inclusion’s sake. I am in the advantageous position of being the sole technical developer of the targeting criteria for the program, so I have a firm understanding on which covariates are essential to avoid confounding.

Therefore, I limited the covariates to only those either directly or implicitly involved in targeting for care coordination outreach. Utilization-based covariates like inpatient use, members spending more than $10,000, and previous emergency utilization are directly used to target members, while covariates like the flag for whether the member was previously a Nationwide Children’s patient are implicitly used. The implicitly-used covariates are ones that are inherent parts of the outreach process. For example, the flag for whether a patient was previously an NCH patient is inherent to the outreach process because we are able to match that member’s eligibility record to their electronic medical record automatically by virtue of having previous billing data at NCH. Members that do not have a medical history at NCH are more likely to require a manually-created medical record. Such members have historically been in a queue that is worked when staffing resources are available, and are consequently at a disadvantage from an outreach perspective.

Another challenge I encountered was the ability to make the document interactive with an end user without hosting the data on a non-NCH server. I solved this by integrating JavaScript code and leveraging some R libraries that allowed me to embed interactive visualizations within an html markdown document. This allowed me to protect the data in accordance with my data use agreement and fulfill the interactive visualization requirement of this project.

Finally, there may be some confounding in my results, but it is difficult to quantify. Although enrollment volume requirements have been recently removed, we were required to manage two percent of the total population for each plan, which amounted to about 2,500 members. I included every known direct and implicit targeting criteria for the care coordination, but a wide net was intentionally cast to meet the aforementioned enrollment requirement. As a result, the population we managed was very diverse. Matching on every nuance in the population would likely cause overfitting in the propensity matching model, but under-fitting the model can cause confounding in my results. To remedy this, I am going to experiment by adding a flag for whether a patient had the top five-to-ten conditions managed in the program. In accordance with state regulations, each enrolled member is required to have a primary diagnosis assigned to him/her. We define that diagnosis as the one that is driving his/her needs for care coordination. This does not account for comorbidities, but may eliminate some confounding in my analysis.