Katherine Ochoa Hernandez

DATA 205 – Capstone

12/14/2023

**Capstone Final Written Report**

**Project Plan**

For my capstone project, I had the privilege to work with Walter Reed Medical Center. As a team, we wanted to 1) discover national trends in annual incidence and annual average cost 2) discover trends in patient demographics 3) examine patient survival 4) discover common associated procedures and 5) run a cost analysis on the procedures. All these tasks will aid in our goal to improve patient care nationwide.

Flap surgeries, also known as free tissue transfer, are a reconstructive surgery where a healthy piece of skin or tissue is transferred from one part of the body to another. These transfers can be composed of skin, fascia, muscle, and bone. Some instances where a flap surgery is performed include trauma related injuries, such as an amputation, or after the removal of a cancerous mass where a flap is used to cover the exposed surgical site. While these procedures come with some risk, there are many benefits associated with them such as stable wound coverage, improved aesthetic and functional outcomes, and minimal donor site morbidities.

The dataset that was used for this project was sourced from the Healthcare Cost and Utilization Project (HCUP) database, which is the nation’s most comprehensive source of hospital care data collected from 1988 - 2020. With an abundance of possible dataset choices, my team decided on the National Inpatient Sample (NIS), which contains data on over seven million hospital stays from all states participating in HCUP, covering 97% of the U.S. population. We ultimately decided on using the 2016 – 2019 datasets. To ensure consistency, we opted to omit datasets prior to 2016 because in 2015 there was a switch from ICD-9 billing codes (internationally used to report diseases) to ICD-10 codes. We also opted to omit 2020 because the pandemic would skew our results since most elective procedures were cancelled.

The dataset includes a wide variety of clinical data, not limited to patient demographic, admission/discharge information, 40 columns of diagnoses, and 25 columns of procedures.

**Data Preparation**

Since patient medical data is very sensitive, the data needed to be purchased through Walter Reed. Once my team accomplished this task, they provided me with the four datasets: Core 2016, Core 2017, Core 2018, and Core 2019. Once I had access to the data, I used R Studio to read in the four individual files. Then, I merged the four files by the common columns. Upon further observation, it became obvious that there were numbers that did not make sense; negative values for binary (0 or 1) categorical variables, ages recorded as negative values, and negative number of days in the hospital. To ensure consistency and reduce skewness, I set all negative numbers equal to NA.

To dig further, I lengthened the data by converting the procedure columns into rows, setting the groundwork to group ICD codes for further analysis. I also filtered out any codes listed as “i10” or “NA” and trimmed any leading or trailing whitespace around the ICD codes. Additionally, I created new columns for the categorical variables with their string equivalencies, a column with the average cost of procedures, and a column with the average number of diagnoses per procedure.

**Descriptive Statistics**

We discovered the following trends:

1. There was an annual increase of flaps procedures at a rate of 1292 surgeries. There was an annual increase of the average cost of procedures at a rate of $6733. (Fig.1)
2. Except for Asian/Pacific Islander, over 50% of patients in each racial group were male. (Fig. 2)
3. The median age of patients throughout all racial groups lies between 50 and 60 years old with White patients on the higher end and Native Americans on the lower end. (Fig. 3)
4. Most of the associated procedures contained the prefix “0JB” which are classified as tissue and fascia excision. (Fig. 4)

A graph with a blue line

Description automatically generated A table of numbers with text

Description automatically generated with medium confidence

Figure 1 Figure 2

A chart of different colored squares

Description automatically generated

Figure 3

A graph of colored bars

Description automatically generated with medium confidence

Figure 4

**Final Data Product**

My final data product composed of multiple resources that medical professionals all around the country can use to improve patient care.

My first two products are general linear models (GLM) that predict future annual incidence and average annual charge of flap surgeries. These models would be beneficial in creating and implementing staff policies, procedures, and budgeting. It would also benefit the public by giving them insight into any potential medical charges they may incur.

Annual Incidence = 1292 \* (year) – 2545754

Average Charge ($) = 6733 \* (year) – 13457644

My third product sought to determine which variables aid in predicting a patient’s likelihood of survival. To accomplish this, I ran a GLM under the binary family using “died” as the response variable with the following predictor variables: year, sex, race, age, length of stay, number of diagnoses, number of procedures, hospital division and household income. Comparing the resulting p-value with an alpha level of 0.05, I concluded that hospital division is not statistically significant and sex is only slightly significant. These results determine that a patient’s chances of survival are not dependent on the hospital division.

My fourth product sought to aid in the decision-making process when a patient is recommended for a flap surgery based on their chances of survival. I ran five individual GLM’s using “died” as the response variable for all and setting the predictor variable to: sex, race, age, hospital division, and household income. The resulting Akaike Information Criterion (AIC) number for each follow: 41804 for sex, 40257 for race, 40385 for age, 41808 for hospital division, and 40919 for household income. Comparing AIC, I concluded that race, with the lowest AIC, best models a patient’s likelihood of survival. These results determine that a patient’s race needs to be taken under consideration when evaluating the necessity for a flaps surgery. Additionally, these results further support the conclusion that hospital division, with the highest AIC, is not an efficient model for patient survival.

My final product is an interactive chart comparing the number of diagnoses for the most expensive primary procedures and their associated average cost. To capture this, I utilized the dplyer package in R where I 1) took the combined dataset (2016-2019); 2) selected the variables number of diagnoses, number of procedures, primary procedure ICD code, and total charge; 3) filtered the observations where the total charge is greater than $900,000 and the number of procedures is equal to 1; 4) grouped by the procedural ICD code; 5) aggregated the following calculations: count of each code, sum of the total charge, average charge per code, sum of the number of diagnoses, and the average number of diagnoses per code. With this new dataframe, I utilized the highcharter package to create an interactive plot with the average number of procedures on the left y-axis, the average cost of a procedure on the right y-axis, and the procedural codes on the x-axis. The bars represent the average number of diagnoses per procedural code and the line represents the average cost per procedural code. We would expect a positive correlation with the average number of diagnoses and the average cost of procedures. However, this is not always the case. The ICD code 0JB70ZZ has the highest average cost but the third least number of diagnoses. Similarly, the code 0HBBXZZ has the greatest number of diagnoses but is on the lower end of the average cost. We can conclude that there are no associations between the number of diagnoses and the average cost of procedures. We can also conclude that the codes with the prefix “0JB”, tissue and fascia excision, tend to be among the most expensive procedural codes with “0HB”, skin excision, coming in second. These results give a decent overall look into the cost of procedures but leave plenty of room for exploration.

The following attachment is a still frame of the final data product.

A graph with red and green bars

Description automatically generated

**Data Story**

My full data story could be found in the following RPubs document: https://rpubs.com/KathyOchoa/1130852

**Overall Experience**

Dataset

* Pro: A lot of data to work with, endless possibilities for analysis, general decent recording of data
* Con: Not all years had the same variables, data is not easily accessible, inconsistency when reporting NA’s

Mentor

* Pro: Very knowledgeable with data science and with its medical application, easy to communicate, multiple available resources
* Con: Very hands off, unclear instructions, busy schedule

**Recommendations**

1. Update instructional video to include R, in addition to SAS
2. Maintain consistency in variable names throughout the years

**Acknowledgements**

Montgomery College:

Rachel Saidi

Abdirisak Mohamed

Michael Iapalucci

Walter Reed Medical Center:

Mentor: Julio A. Rivera, Ph.D.

Team: Colin J. Harrington, M.D.

Jordan Troph, M.D.

Intern Coordinator: Toby Perkins

Personal Support:

Steven Guerrero

Family

DATA 205 Classmates

Sources

* HCUP
* MediVisuals
* Dr. Tjoson Tjoa UCI Head & Neck
* Mayo Clinic
* American Cancer Society
* ICD10Data.com
* GeeksForGeeks.org