An Analysis of Changing Prices With Respect To Illness Severity and Ethnicity From 2017 to 2022

Fasih Munir
Computer Science Department
City, University of London
London, United Kingdom
fasih.munir@city.ac.uk

***Abstract* — In this paper we will present an analysis specific to New York City, USA, looking at data from 2017 to 2022. We will look to understand if prices for healthcare have significantly increased and if certain groups of people are affected by this. As we explore the data we will find that there is a significant difference between prices with high dispersion at the minor illness level. Further exploration will reveal that multi-ethnic groups for minor illnesses have the highest variability for their cost of care and tend to cluster near the Manhatten area in New York.**

I. Introduction

Given a similar illness severity categorization (described by the All Patient Refined Diagnosis Related Groups Methodology [1] as the extent to which your body deteriorates and loses function, coupled with co-morbidities, ranging from class 1 to 4 where 1 is minor and 4 is extreme) amongst groups of people, would it be reasonable to expect one group to end up paying more for their treatment? There has been research done suggesting that being a part of some ethnic groups can lead to different susceptibilities to diseases [2], which is to say that it could be reasonable to expect groups to pay differently for similar classifications. Some people might have a minor classification for skin problems and pay less than others for the same illness due to the way their genes reacted to the disease making it easier to treat. However, knowing that such differences can exist and with rising costs around the world, it becomes important to try and understand how costs vary between illness severities and how they affect different ethnic groups so that steps can be taken to help manage patient outcomes better.

II. Analytical Questions

In this paper we will aim to dive into 3 questions that funnel into each other.

We will first begin by exploring if in New York City, over the most recent 6 years (2017 to 2022), it has become more expensive to become sick given the 4 groupings of illness severity. If prices have not increased then we can end our investigation, but if prices have risen and they have risen in a manner where the differences are statistically significant between groups and years, then we have a case for continuing our exploration.

We will then move onto understanding how these rising costs affect certain groups of people. We will especially look at ethnicity, but also consider the impact of pricing on other demographic characteristics such as gender and race.

Finally, given our findings from the previous question, we will aim to see which counties of New York City (where the hospitals are located) are receiving the most burden of the identified impacted groups.

The 3 questions together will help us understand the impact of rising costs on certain groups of people that can then help implement targeted strategies to manage patient outcomes

III. Data (Materials)

The data is freely available online [3]. Specifically we are working with data for New York City, USA. There are 6 datasets for the years 2017 to 2022 for “Hospital Inpatient Discharges (SPARCS De-Identified)”. There are approximately 13 million observations and each year has approximately 2 million observations. The sample is large enough for us to be confident in the results we will obtain from significance testing and clustering.

There are 33 columns and they provide information on where the treatments were administered, de-identified demographic information of the patient, illness categorization and assessment of the patient, payment methods and associated costs. Each observation refers to a discharge, which means that we are looking at observations of patients that were admitted, treated and then allowed to go home. This means that we do not know if there are returning patients, so we will assume each discharge to be a new patient.

This data is supplemented by latitude and longitude points for hospital counties scraped using geopy. This data will enable us to visualize our locations to help answer question 3.

There are certain limitations to keep in mind. The data is USA centric and specifically focused on New York City which makes generalizability to other areas difficult. Since the data is de-identified, there are NaN values which are expected due to purposeful redaction, for example not all hospital counties were scraped for location. Inclusion of timestamps for patient admission could have added additional elements to observe trends. Outliers have been retained as we are assuming them to not be mistaken entries but actual data points. Lastly, some of the demographic characteristics could have been more robustly captured.

IV. Analysis

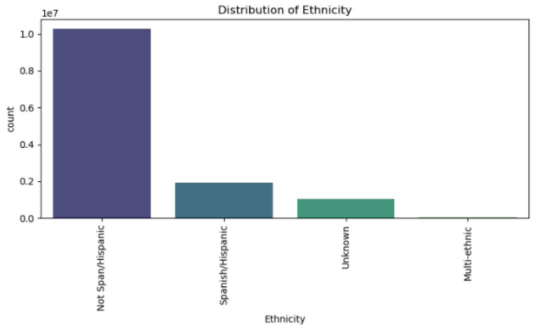
Our approach is as follows:

1. We will first begin by doing some pre-processing

To start, we drop redundant columns, check for duplicates and fix certain strings to proper case. The main action however is scraping locations of the hospital counties. Since question 3 involves observing which areas in New York City have a deep concentration of impacted groups, we would need a way to plot these locations on a map. We use the python client geopy to achieve this. Three inputs are required to be able to get the latitude and longitude points; location name, city, country. The location name is provided in our dataset while the city and country we know to be New York City, USA. The locations are saved and merged back onto our main dataframe using a left join.

2. We then do some initial exploration of the data

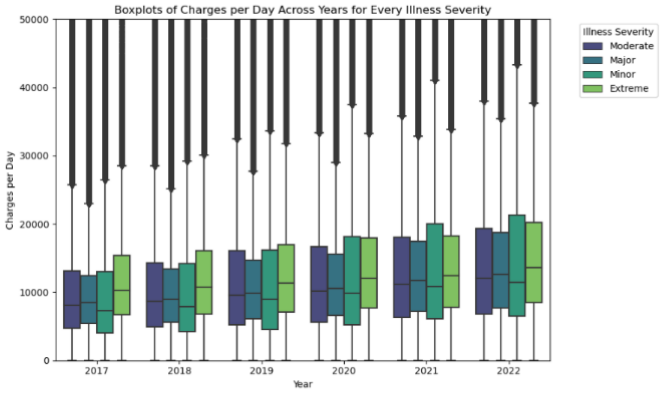
Since most of our data is categorical, we look at their distributions using countplots. None of the variables look to follow the normal distribution and many are right skewed such as "Hospital County", the risk of mortality and ethnicity.



3. We then make our exploration more specific and focus on the prices charged by hospitals

Before diving deeper, we first create a new variable called "Charges per Day" by dividing the total charges with the length of stay. We do this because the data natively gives us the total charged by the hospital for a discharge, however this can create a very large scale making comparisons difficult. We will use the Charges per Day going forward.

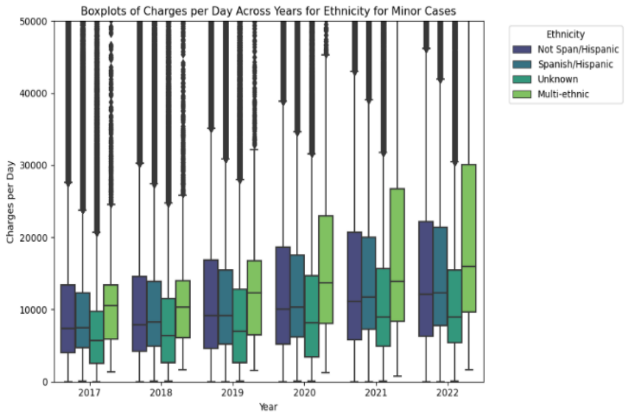
Next, we create a boxplot to see the dispersion of Charges per Day against the Severity of Illness across all the years in our data set.



There is a significant presence of outliers and while the median for the minor illness has always been lower, we do observe an increase in the dispersion over time eventually leading to the minor illness upper quartile to be higher than all the rest. This could mean that there are more unusual cases coming in the minor category that require more specialist treatment. Significance tests (Welch test, assuming unequal variance which can be seen in the box plots) at the 5% level comparing the averages for every combination of groups for every year were also run returning significance for all combinations bar one meaning that the variations seen are not due to random chance. It has become more expensive to be sick and if you have a minor illness you might be charged a higher amount than other groups.

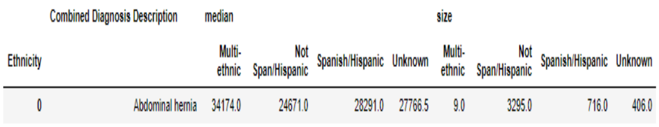
4. After setting the stage for the increased prices, we continue exploring the effect on individuals

We follow a similar approach to the previous section except this time we drill deeper into minor illnesses only since it has the most variability in recent years.



The dispersion for the multi-ethnic group is more dramatic compared to the rest. They have always paid a higher median value and their variability has also increased. We might be seeing this variability due to the lower sample sizes for this group across years (approximately 1200 observations each) however, building similar significance tests like before (5% level), for all combinations except 2, the differences are significant and thus not by chance.

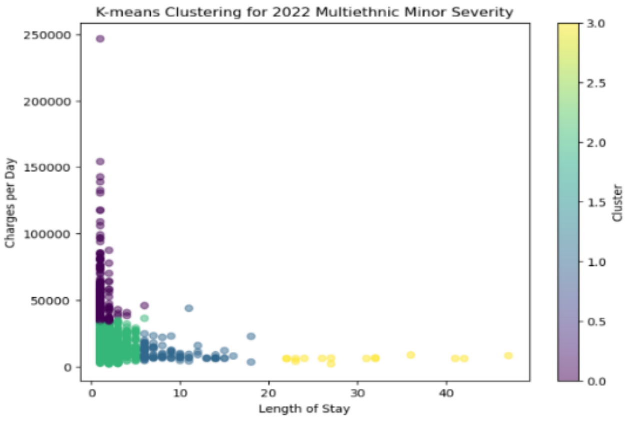
Tying this back to our initial research aim, to see if ethnic differences result in different charges, it can be the case that people from a multi-ethnic back ground react differently to illnesses for example an Abdominal Hernia where they paid the highest charge.



	Combined Diagnosis Description	median				size			
Ethnicity		Multi-ethnic	Not Span/Hispanic	Spanish/Hispanic	Unknown	Multi-ethnic	Not Span/Hispanic	Spanish/Hispanic	Unknown
0	Abdominal hernia	34174.0	24671.0	28291.0	27766.5	9.0	3295.0	716.0	406.0

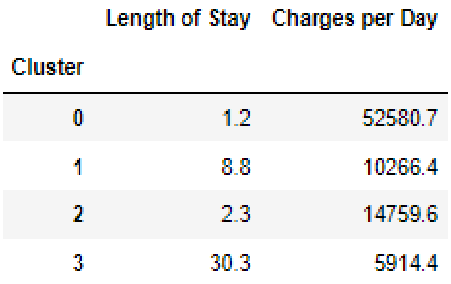
Other researchers have arrived at similar conclusions for example non-hispanic black patients paying more for skin ulcer treatment compared to other groups [3].

To understand the sub groups that form within the multi-ethnic people, we create a K Means clustering model. We use the most recent year of 2022 as it is assumed that it will have the most up to date health care policies and interventions (cannot use outdated policies as a reason for differences). Since K Means clustering is a distance based method (distance to nearest centroid) we scale our filtered data which includes the daily charge and length of stay. Using an elbow plot we find our optimal number of clusters to be 4 resulting in the following plot:



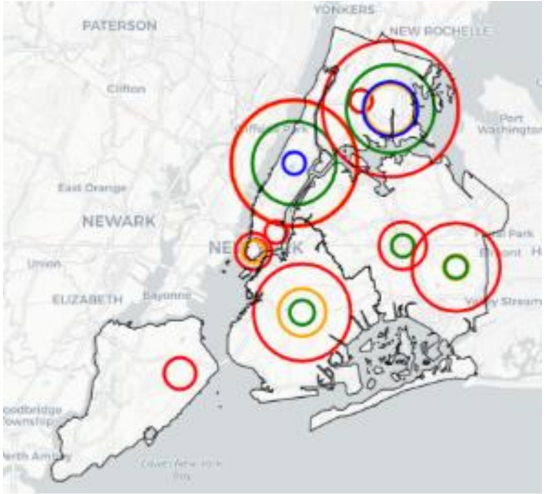
Cluster 2 (green) has a little bleed with Cluster 0 (purple) and Cluster 1 (blue). As length of stay decreases the charges per day increase dramatically.

The number of patients discharged in Cluster 3 (yellow) are low. Most discharged patients come for shorter durations (Clusters 0 and 2). They likely come with specific complications that increase the cost of their treatment.



Cluster	Length of Stay	Charges per Day
0	1.2	52580.7
1	8.8	10266.4
2	2.3	14759.6
3	30.3	5914.4

5. Finally, we look at how the clusters manifest within New York City

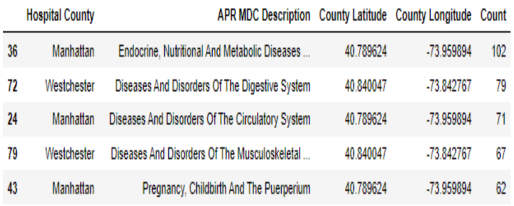


We use folium to generate the above map. The larger rings represent weight or the count of observations. The cluster colours are as follows: (cluster, length of stay, daily charges)

0, 1.2, 52580.7 = orange
1, 8.8, 10266.4 = green
2, 2.3, 14759.6 = red
3, 30.3, 5914.4 = blue

The counties of Manhatten and Westchester have the highest concentration.

This does not necessarily mean that people live in these areas but it means that they had to go to the hospitals in these counties to get their treatment. Either because these were the only places where they were available or because these were the closest areas at the time of illness.



	Hospital County	APR MDC Description	County Latitude	County Longitude	Count
36	Manhattan	Endocrine, Nutritional And Metabolic Diseases ...	40.789624	-73.959894	102
72	Westchester	Diseases And Disorders Of The Digestive System	40.840047	-73.842767	79
24	Manhattan	Diseases And Disorders Of The Circulatory System	40.789624	-73.959894	71
79	Westchester	Diseases And Disorders Of The Musculoskeletal ...	40.840047	-73.842767	67
43	Manhattan	Pregnancy, Childbirth And The Puerperium	40.789624	-73.959894	62

The above table shows the top 5 counties with respect to the highest count for types of illnesses.

V. Findings, Reflections and Further Work

Based on the findings, we could recommend to the city of New York to pay special attention to their multi-ethnic people near Manhatten and Westchester and provide them with cheaper access to treatments for minor severity metabolic diseases. While this is useful for the city to know, it is difficult to recommend something immediately actionable. Further research around the kinds of diseases that are faced by specific groups would be required to make a robust recommendation. We could look at specifically which diseases are affecting these people, what equipment is required, could this be acquired in bulk to reduce cost and can we upskill doctors to treat outlier cases.

It is also important to note that the ethnicity variable itself is fuzzy. It contains 4 categories, Hispanic, not Hispanic, multi ethnic and unknown. There are many more ethnicities that exist outside of these 4 and these alone do not specifically capture a kind of group. They are quite open ended categories, which makes it hard to provide a concrete plan for a group of people.

While the data has been suitable as a starting point, further refinement in the collection process and subsequently the data preparation process is required. Given these changes regression models could be created to better understand causality.

The distributions that were checked independently could be checked as a combination as well to see how connected features interact for example female_hispanic rather than female separate and Hispanic separate.

Separate analysis could also be done on outliers to get a better understanding of where the variability is coming from and if the data could be grouped better to remove this.

This analysis went down the minor cases for certain groups of people. Similar work could also be done for the other severity groups to give a more holistic picture about the health outcomes.

References

[1] R. Averill et al., “ALL PATIENT REFINED DIAGNOSIS RELATED GROUPS (APR-DRGs) Methodology Overview 3M Health Information Systems,” 2023. Available: https://hcup-us.ahrq.gov/db/nation/nis/APR-DRGsV20MethodologyOverviewandBibliography.pdf

[2] R. A. Bulatao, N. B. Anderson, and in Later, “The Nature of Racial and Ethnic Differences,” Nih.gov, 2015. https://www.ncbi.nlm.nih.gov/books/NBK24684/

[3] “State of New York | Open Data Health | State of New York,” New York State Department of Health | Health Data NY. http://www.health.data.ny.gov/

[4] N. J. Hardy, C. Gronbeck, and H. Feng, “Impact of race and ethnicity on length of stay, discharge location, and total charges for inpatients with skin ulcers in New York,” Archives of Dermatological Research, vol. 315, no. 7, pp. 2187–2189, Apr. 2023, doi: https://doi.org/10.1007/s00403-023-02624-3

Word Counts:

1. Abstract – 97 words
2. Introduction – 187 words
3. Analytical Questions – 193 words
4. Data (Materials) – 276 words
5. Analysis – 911 words
6. Findings, Reflections and Further Work – 305 words