LGBTQ+ Community Health Status in the United States

Kari Lauro

SI 618, Data Gathering and Processing Report

Motivation

Access to quality healthcare is an essential aspect of life, and unfortunately, many individuals don't seek care due to cost, lack of representation, stigma, discrimination, etc. (Kates, 2018). Though many studies have determined the main contributor to avoidance of care for the LGBTQ+ community is provider discriminatory beliefs, there is still a vast opportunity for further search within this area to determine additional measures that can be taken to address and further reduce disparities that exist in healthcare for this population (Aleshire et al., 2018).

The initial set of research questions surrounded diversity within the different aspects of healthcare services/service areas. Upon starting the data analysis process of each of the datasets listed below, these questions were changed as follows:

Questions

1. Are health policy protections and/or discriminations related to the perception of health concerns being taken seriously by providers?
2. How do perceptions of stress differ for overall health compare to perceptions of stress specifically for the LGBTQ+?
3. Is the stress from being part of the LGBTQ+ community related to a lower perception of medical concerns taken seriously?

Data Sources

To gain insights into the health of the LGBTQ+ community, data was gathered from a total of three datasets.

Mapping LGBTQ Equality: 2010 to 2020, United States

Retrieved from the Institute for Social Research in partnership with the University of Michigan with access here: https://www.icpsr.umich.edu/web/RCMD/studies/37877. While in the initial proposal, my intention was to download this as a SAS file, I found the process of reading the data too complicated and instead exported it in a delimited format as a .tsv file. With 59 total rows, this data included "tallied" data regarding policies that either positively or negatively impacted the LGBTQ+ community. Each positive policy was added as +1, and each negative policy was subtracted as -1, with higher tallies being associated with a greater number of positive policies. Relevant columns included Census four region, 2020 overall healthcare tally, and Number of state adults who are LGBT.

National Couples' Health and Time Study (NCHAT) 2020
Data was retrieved from the Institute for Social Research, accessible here: https://www.icpsr.umich.edu/web/DSDR/studies/38417.
This data was also exported in a delimited format as a .tsv file and contained a total of 3,639 rows. Included in this dataset were a variety of health survey responses. Relevant columns included Age on survey date (in years), What sex appears on original birth certificate, People who are LGBTQ+, In general, would you say your health is, etc. Though most of the data were available for public use, some were noted as restricted. Unfortunately, my request to access this data was denied, and I did not get to use the state/region data that I was hoping for.

The US Health Insurance Dataset
Retrieved from Kaggle and can be accessed here: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/versions/1?resource=download.
Exported in a .csv format and containing 1,338 rows, this insurance focused dataset, contains information about insurance charges in relation to age, sex, BMI, number of children, smoking status, and region for the year 2020. This dataset provided the additional link needed to join my first two datasets together.

Data Manipulation Methods
While each of the datasets contained information necessary to complete analyses for each of the three research questions, it also contained extraneous data that would complicate this process. This section reviews the process I used read in each of the datasets and filter the datasets down to the components needed for analysis.

Reading in the data through Spark
Each of the datasets was read using *spark.read.option* to ensure each would be read in properly and had the appropriate headings, before being subsequently assigned to their own variable. Once properly read-in, each dataframe was converted into a table using *.registerTempTable* before moving forward to the next process of refining and joining the tables for further analysis.

Filtering and joining the datasets
In the process of cleaning the datasets to allow for appropriate analysis, I first needed to filter each of the datasets down to remove extraneous data. To capture the appropriate columns I used sparkSQL for both Mapping LGBTQ Equality and NCHAT. From here, the filtering and process of the data began. When needed, I replaced values from NCHAT that were numeric with their representative description using *regexreplace()*. Once satisfied that I had what I needed, it was time to join each of my three tables together. Keeping in mind the size difference between them I used SQL queries to join the tables using the *BROADCAST()* clause to help avoid repeated data and skewness that would be caused with a regular *JOIN*. Further filtering out of null/missing data was addressed per research question if needed using *df.filter()* to remove rows with null values.

The final table consisting of the data used to answer each research question is found below in Figure 1.

sex	region	charges	2020_health	No_LGBT	Age	LGBTQ_stress	relationship_stat	health_stat	insurance	health_identity	trust_provider	health_seriously
male	northeast	6406	6.0	26000	1	3	2	4	3	3	3	4
male	northeast	6406	6.0	26000	2	5	1	3	1	3	4	2
male	northeast	6406	6.0	26000	2	4	2	5	1	1	4	2
male	northeast	6406	6.0	26000	1	5	1	2	4	5	1	4
male	northeast	6406	6.0	26000	2	4	1	1	1	4	4	2

Figure 1: Final table used for analysis

Analysis and Visualization

Q1. Are LGBTQ+ health policy protections and/or discriminations related to the perception of health concerns being taken seriously by providers?

To properly answer this question, the three data sets were queried multiple times to filter down the relevant information needed and to remove unnecessary data. A final query was used to compare whether a higher tally of health policies would impact whether or not an individual felt their health concerns would be treated seriously by healthcare providers.

```
"SELECT 2020_health AS num_health_policies,
ROUND(AVG(CAST(health_seriously as int)),2) AS
avg_perception_health_taken_seriously FROM full_data WHERE
health_seriously !=-98 GROUP BY 2020_health ORDER BY 2020_health DESC"
```

As seen in Figure 2, interestingly, it appears as though concern over whether or not health concerns would be taken seriously was not impacted by a higher or lower total tally of health policies. The average ranking of concern was the same across each documented total tally of health policies across the United States. With concern that this may be due to a calculation error, I attempted to re-filter both the tally of health policies and the levels of agree with concern over whether health concerns would be take seriously as well as changing both my *GROUP BY* and *ORDER BY* statements but received the same average level of agreement per tallied health policy each time.

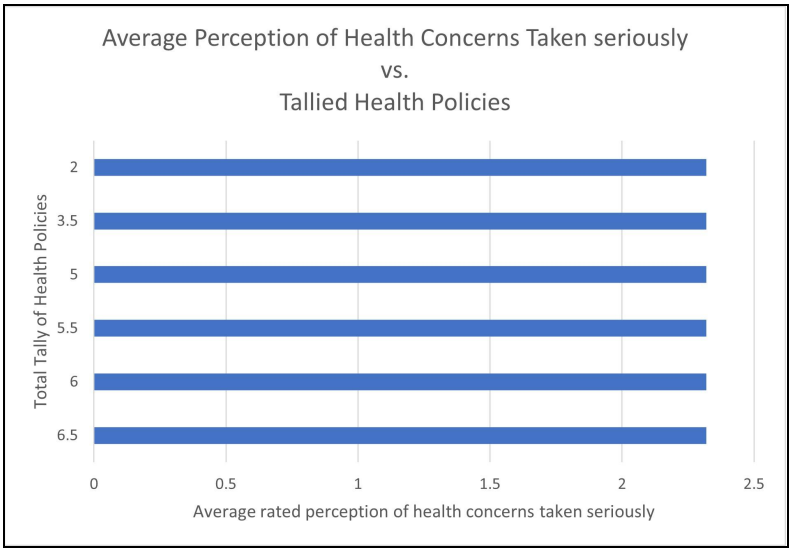

Figure 2: Bar graph showing the perception of the likelihood of health concerns being taken seriously compared to the different number of tallied LGBTQ+ health policies

Q2. How do perceptions of stress compare for overall health for the LGBTQ+ Community?

In analyzing this question, the three datasets were queried multiple times to narrow in on the health-specific data. From this more refined, health-focused dataset, another query was performed focusing on comparing rated LGBTQ+ stress to rated health status. To perform this analysis, an average rating of LGBTQ+ stress was calculated and grouped by the different ratings for health status.

```
"SELECT health_stat AS perceived_health_status,
ROUND(AVG(CAST(LGBTQ_stress as int)),1) AS avg_lgbt_stress FROM
perceived_health WHERE LGBTQ_stress != 99 GROUP BY health_stat ORDER BY
avg_lgbt_stress DESC"
```

Unlike question #1, there does appear to be a trend between stress felt by members of the lgbtq+ community and health status. Figure 3 shows that as health status improves stress decreases. Also of note, stress was highest in those who refused to answer their health status on the survey. Further analysis could be key here in determining what exactly the relationship is between the declination to respond about health status and stress felt by the LGBTQ+ community.

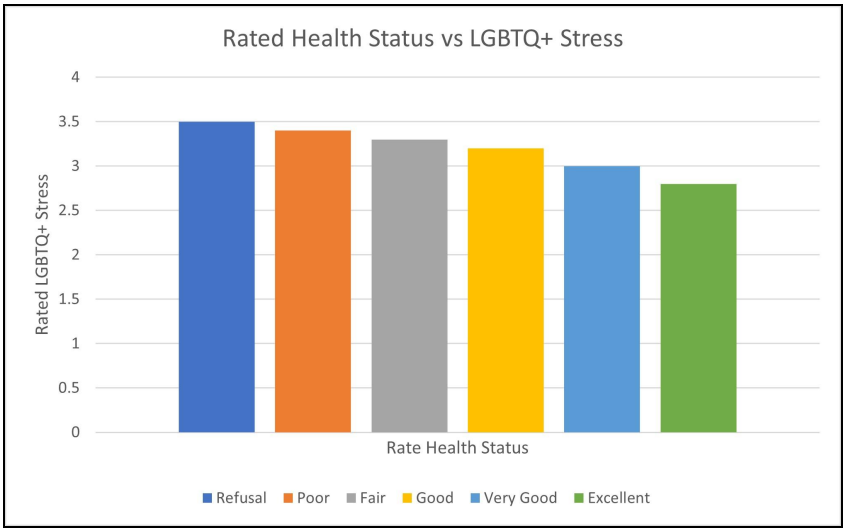

Figure 3: Bar Chart showing average rating of stress for members of the LGBTQ+ community compared to health status

Q3. Is the stress from being part of the LGBTQ+ community related to a lower perception of medical concerns taken seriously?

Finally, to answer question #3, the process started with the already refined, health-focused datasets from questions #1 & #2. From here, a query was devised to determine if there was any relationship between whether or not individuals felt that their health concerns would be taken seriously compared to the stress felt by the lgbtq+ community.

```
"SELECT health_seriously AS Concern_healthcare_due_to_identity,
ROUND(AVG(CAST(LGBTQ_stress as int)),1) AS avg_lgbt_stress FROM
perceived_health WHERE health_seriously !=-98 GROUP BY
Concern_healthcare_due_to_identity ORDER BY avg_lgbt_stress DESC"
```

While both stress related to being a member of the lgbtq+ community and concern about whether or not health concerns would be taken seriously by medical providers both had a decreasing trend line, when comparing the two values with each other, there appears to be no correlation between them. This can be visualized in Figure 4. I did try to conduct a further analysis on this question to see if there was any relationship between either of these variables with overall health status but could not get the proper grouping I needed to clearing see any patterns at this time.

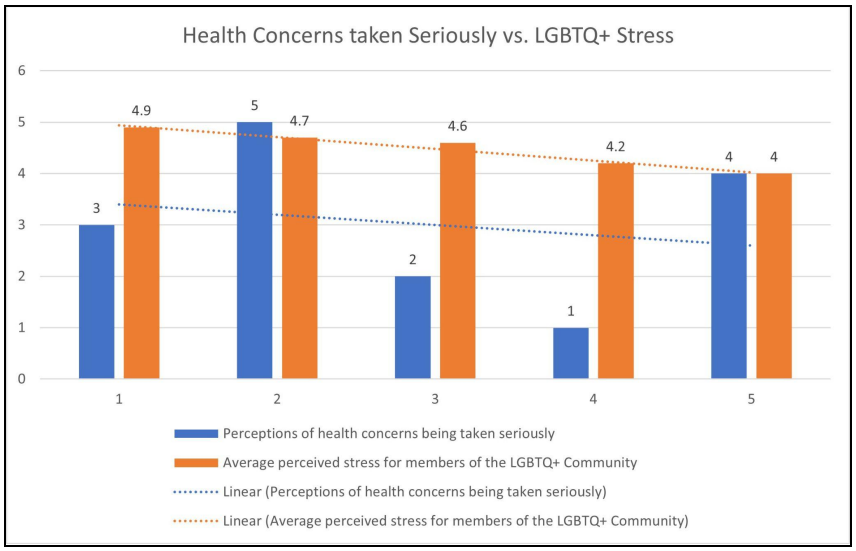

Figure 4: Bar chart illustrating stress for the LGBTQ+ community and whether individuals felt their health concerns were taken seriously

Challenges

Upon starting the cleaning of the NCHAT dataset, it was uncovered that the columns of data that would've ideally been used to perform joins between my initial two tables were restricted and though I did request access, it was not granted and I needed another method to properly join my tables without risking too much skew. Additionally, one I added the health insurance table, I was able to join my data but quickly realized that I wanted to shift the focus of my original research questions and thus had to alter which pieces of data I was utilizing from each dataset. Finally, though I had planned on using spark *map()*, *reduceByKey()*, and *sortByKey()* to analyze my data, I felt I could get a better understanding by using sparkSQl and therefore shifted my final analyses to using queries to filter down, group, and perform calculations on my data.

Resources

Aleshire, M. E., Ashford, K., Fallin-Bennett, A., & Hatcher, J. (2018). Primary care providers' attitudes related to LGBTQ PEOPLE: A narrative literature review. *Health Promotion Practice*, *20*(2), 173–187. https://doi.org/10.1177/1524839918778835

Jennifer Kates (2018, May 3). *Health and access to care and coverage for lesbian, gay, bisexual, and transgender (LGBT) individuals in the U.S.* KFF. Retrieved October 6, 2022, from https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-and-access-to-care-and-coverage-for-lesbian-gay-bisexual-and-transgender-individuals-in-the-u-s/

Movement Advancement Project. Mapping LGBTQ Equality: 2010 to 2020, United States. Inter-university Consortium for Political and Social Research [distributor], 2021-07-14. https://doi.org/10.3886/ICPSR37877.v2

Kamp Dush, Claire M., and Manning, Wendy D. National Couples' Health and Time Study (NCHAT), United States, 2020-2021. Inter-university Consortium for Political and Social Research [distributor], 2022-07-14. https://doi.org/10.3886/ICPSR38417.v1

Datta, A. (2020, February 16). US Health Insurance Dataset. Kaggle. Retrieved October 27, 2022, from https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download