**About Our Dataset**

Hospital Compare is a website operated by the Centers for Medicare and Medicaid Services (CMS) under the Department of Health and Human Services (DHS) by the United States federal government. The website provides access to official datasets collected from over 4,000 Medicare-certified hospitals across the country as part of “an Administration-wide effort to increase the availability and accessibility of information on quality, utilization and costs for effective, informed decision-making”.[[1]](#footnote-0) The information is intended to allow consumers to select and directly compare performance measures of various hospitals, and thereby empower consumers to make informed decisions regarding their health care.

The Hospital Compare website (made available online at: <https://data.medicare.gov/data/hospital-compare#>) contains 58 separate datasets which include various performance-related measures, including records of complications, hospital-associated infections, and even patient-satisfaction survey data. Many of the datasets are provided at the national, state, and individual hospital level so that researchers/consumers can reference national and statewide performance as well as the performance of specific facilities. The hospital-level tables tend to be the largest in size, usually containing around 70,000 records.

According to the Hospital Compare website, all datasets are publicly accessible and exist in the public domain. Within the documentation associated with the Hospital Compare dataset, the following statement can be found:

All Hospital Compare websites are publically [sic] accessible. As works of the U.S. government, Hospital Compare data are in the public domain and permission is not required to reuse them. An attribution to the agency as the source is appreciated. Your materials, however, should not give the false impression of government endorsement of your commercial products or services.

So long as researchers/consumers provide attribution and do not claim any government endorsement of a developed product, use of Hospital Compare data is fairly unrestricted.

Complementing the dataset, CMS provides a “Hospital Compare Downloadable Database Data Dictionary” to help researchers/consumers interpret undefined acronyms and terms used throughout the 58 datasets. The explicit purpose of this dictionary is “to provide a directory of material for use in the navigation of information contained within the Hospital Compare downloadable databases”.[[2]](#footnote-1) While this documentation does assist with interpreting the definitions of various terms, there are still many questions involving methodology that are left unanswered. For example, many of the databases use scoring metrics to rate quality of care at various hospitals; however, the CMS documentation fails to explain the criteria for how these scores were decided. As a result, researchers/consumers may be reluctant to make use of certain metrics or dataset for fear of misinterpreting data.

**Cleaning Process**

After selecting the Hospital Compare dataset for our project, our group agreed to the following research question:

*Some hospitals claim to have better service and charge more for high quality care, but are patients really getting their money's worth?*

In order to answer this question, we need to evaluate and compare: (1) the quality of care being received by patients at Medicare-certified hospitals; and (2) the average spending per patient for procedures. Our group plans to use HCAHPS hospital survey data and spending data, both provided in the Hospital Compare database. For the purposes of our project, our group decided to limit our scope to consideration of Medicare-certified hospitals in the DMV (DC-Maryland-Virginia) area.

To clean this data, we identified our first step: to remove all non-relevant states from all datasets. Since we only intended to compare hospitals in the DMV area, we filtered out all non-DMV states; however, after doing so, we realized that **CMS had suppressed all Maryland hospital data**. This left our group with only DC and Virginia hospital data to analyze and compare. Second, we discussed how to remediate any missing data. While our datasets do not contain any empty cells, many cells were left as “Not Applicable” or “Not Available.” Initially we thought to remove this data, but after discussing that these entries might be useful for interpreting our survey response data, we decided to keep it, at least for now. After taking these steps, we are left with cleaned patient-satisfaction survey data and patient spending data. Currently, our group is unsure how to join these two very different datasets.

To conclude, we narrowed our dataset to contain only information relevant to our research question; however, there are still some inconsistencies. We hope to receive advice on how to approach these inconsistencies/barriers before moving forward with further cleaning.

Word Count: 736

1. *Hospital Compare Downloadable Database Data Dictionary* (pp. 1-81, Publication). (October 2016). Accessed on 10/28/2016 at:

   https://data.medicare.gov/views/bg9k-emty/files/5440a626-3030-41cc-8531-8aae0b3902ef?filename=DDB\_Data\_Dictionary\_October\_2016\_Final.pdf&content\_type=application%2Fpdf%3B%20charset%3Dbinary [↑](#footnote-ref-0)
2. *Hospital Compare Downloadable Database Data Dictionary* (pp. 1-81, Publication). (October 2016). Accessed on 10/28/2016 at:

   https://data.medicare.gov/views/bg9k-emty/files/5440a626-3030-41cc-8531-8aae0b3902ef?filename=DDB\_Data\_Dictionary\_October\_2016\_Final.pdf&content\_type=application%2Fpdf%3B%20charset%3Dbinary [↑](#footnote-ref-1)