Healthcare Data Mining and Cleansing: A Study of Improving Data Quality for Effective Data Analysis on NPPES

Author: Carlos A Pérez Medina
Advisor: Dr. Nelliud Torres
Master in Computer Science

Abstract

This article examines the use of data mining in the healthcare industry, with a particular emphasis on best practices for increasing data quality, preserving provider information, and applying advanced techniques to extract valuable insights from complicated data sets. The research gathered information from the National Plan and Provider Enumeration System (NPPES) and then analyzed the data to determine whether or not there were any problems with the information. The information was sorted into its two basic groups, which were establishments and service providers. The headers were modified accordingly, the information was standardized by the application of analysis and processing, and any null values have been removed. In the context of the data utilization on healthcare provision across the nation, questions of ethics, including the protection of individuals' right to privacy and the confidentiality of health information, were discussed and highlighted as critical components

Introduction

The National Plan and Provider Enumeration System (NPPES) is a database maintained by the Centers for Medicare and Medicaid Services (CMS) that includes information about healthcare providers in the United States. The NPPES database includes information about a wide range of healthcare providers, including physicians, dentists, nurse practitioners, and other healthcare professionals.

Background

The NPPES database is a valuable resource for researchers, policymakers, and healthcare professionals who are interested in understanding the healthcare system in the United States. The database includes a wealth of information about healthcare providers, including their specialties, practice locations, and contact information.

Overall, data analysis of the NPPES database has the potential to provide valuable insights into a wide range of healthcare issues. By using this database to study healthcare access, utilization, and outcomes, researchers and policymakers can develop strategies to improve the healthcare system in the United States and to ensure that patients receive high-quality, cost-effective care.

Problem

The main problems associated with the NPPES file is the accuracy and completeness of the data it contains. While the NPPES database is intended to be a comprehensive and up-to-date record of all healthcare providers in the United States, there are many factors that can affect the quality of the data it contains. For example, providers may change their practice locations or contact information without updating their records in the database, or may fail to provide accurate information about their specialties or credentials. In addition, there may be errors or inconsistencies in the way that data is entered into the database, which can make it difficult to use the data for research or policy purposes. As a result, researchers and policymakers who use the NPPES database must take care to validate the data they are using and to account for any limitations or biases in the data.

Methodology

Data Collection:
The data used in this study was obtained from the National Plan and Provider Enumeration System (NPPES), which can be found at https://download.cms.gov/nppes/NPI_Files.html, which is maintained by the Centers for Medicare & Medicaid Services (CMS). The NPPES database contains information on healthcare providers and suppliers, including demographic information, specialty information, and practice locations. The information was received as a CSV file in a safe, central location.
Data Quality
A data quality assessment was conducted to identify any issues or anomalies in the data. This process involved a review of the data to identify missing or inconsistent values, duplicate records, and other potential issues. A first look was done using a sample of about 100 records that were chosen to be representative. Upon examination of the file, it was observed that a significant number of records contained null values, which could potentially impact the validity and accuracy of the overall analysis. The file comprises 330 column headers, making it difficult to apply conventional analytical methods. Upon loading the file into a Python environment, it was determined that the file contained approximately 7,436,413 records and had a size of approximately 9 GB. The file was downloaded in comma-separated value (CSV) format, which can present significant challenges for individuals without extensive experience in handling large data sets.
Data Analysis
The subsequent step in the process was to conduct a comprehensive data analysis. To commence this process, the data was initially analyzed to determine the format in which it was provided. Upon reviewing the file, it was determined that the information was divided into two primary categories, referred to as "entity types." These entity types were identified by "1" or "0" values As shown in Table 1, which indicated whether the information pertained to a facility or a provider.
Data Quality:
The subsequent step in the process was to conduct a comprehensive data analysis. To commence this process, the data was initially analyzed to determine the format in which it was provided. Upon reviewing the file, it was determined that the information was divided into two primary categories, referred to as "entity types." These entity types were identified by "1" or "0" values As shown in Table 1, which indicated whether the information pertained to a facility or a provider.2 (Facility) 1,038,038

Entity Type	NPI Count
1 (Providers)	5,639,171
2 (Facility)	1,038,038

performed, and no duplicates were found. The null values present in the 330 columns were then removed, and a dictionary was created as a reference table to standardize the information pertaining to the providers and facilities. This allowed for a well-organized and manageable dataset that could be easily analyzed. It was observed that the columns were not in their correct order. Subsequent adjustments were made to the data, making it ready for a more formal analysis. The first analysis performed was the validation of provider names, which contained dirty data, including non-alphanumeric values such as "","","","," and others.

Results and Discussion

The accuracy and completeness of the data contained within the NPPES database is a critical factor in its usefulness for research and policy purposes. The steps taken to clean and validate the data for our analysis.

To ensure the accuracy and completeness of the data used in our analysis, we undertook a rigorous data cleaning process. The data cleaning process involved several steps, including:

Identifying and removing duplicate records: We used the provider's National Provider Identifier (NPI) number to identify and remove any duplicate records from the dataset.

Validating practice location data: We validated practice location data by cross-referencing it with external datasets.

Standardizing provider specialties: We standardized provider specialties by mapping them to a set of pre-defined categories to ensure consistency across the dataset.

Removing incomplete or invalid data: We removed any incomplete or invalid data from the dataset, such as missing contact information or incorrect NPI numbers.
Validating data using statistical methods: We used statistical methods, such as outlier detection and frequency analysis, to identify and remove any data that was outside of expected ranges or appeared to be anomalous. Table 2 shows the top state with providers after data cleanup

State	Number of providers
California	866,083
New York	545,129
Florida	484,922
Texas	466,696
Ohio	295,998

While the NPPES database contains a wealth of information about healthcare providers, the quality of the data can be variable. As a result, it is critical that any analysis of the NPPES data be preceded by a thorough data cleaning process to ensure the accuracy and completeness of the data. This will enable researchers and policymakers to make informed decisions based on reliable data, rather than being hampered by inaccurate or incomplete information.

In conclusion, the NPPES database is a valuable resource for researchers and policymakers interested in understanding the healthcare system in the United States. However, the accuracy and completeness of the data contained within the database can be variable, making a rigorous data cleaning process critical for any analysis. By undertaking a thorough data cleaning process, we were able to generate a reliable and clean dataset that was suitable for research and policy purposes.

Conclusions

The healthcare industry is facing a significant challenge in effectively managing the large volume of data produced daily. To address this issue, the establishment of a standardized information database for service providers is crucial. This reference table serves as a foundation for ensuring data accuracy and consistency, thereby enabling analysts to focus on more important validation tasks, such as claim data and patient information which are subject to frequent changes. Having a centralized and standardized database for service providers will provide a complete overview of a provider and minimize the need for manual searching or contacting the provider or facility for exact information. This will also eliminate inconsistencies in spelling the provider's name and other data, thereby promoting a standardized approach to data management and analysis in the healthcare industry.

Future Work

The prospective work for this study involves the development of a comprehensive and precise reference table that will include all the excluded information due to outdated or inaccurate data. The proposed reference table will be presented to insurance companies to aid in creating a standardized structure, which will assist healthcare analysts in accessing the provider and facility information in a consistent and normalized manner. This will facilitate the generation of reports by decision-makers, without the need for data analysts to cleanse or normalize the data. The ultimate goal of this effort is to streamline data analysis and reporting, allowing decision-makers to focus on their core responsibilities and make more informed decisions, without the burden of managing data quality.

Acknowledgements

I would like to express my deep gratitude and appreciation to Dr. Nelliud Torres, Ing. Edwin Rodriguez, and Data Analyst Jose Vega for their invaluable support and guidance throughout the course of my research. Their insights, feedback, and expertise were instrumental in helping me to navigate the complex landscape of healthcare data analysis.

References

Wang, R. Y., & Strong, D. M. (1996). Data quality research: past, present, and future. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, ACM.

Health Information Management Systems Society (HIMSS). (2018). Electronic Health Records Data Quality, HIMSS.

Liu, B., & Luo, X. (2010). Data quality: concepts, methodologies and techniques. Springer.

[4] Wirth, A., & Hipp, J. (2007). Data preprocessing in data mining. In Data Mining, Springer, Berlin, sHeidelberg, pp. 61-86.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. Pearson Education India.

Han, J., & Kamber, M. (2006). Data mining: concepts and techniques. Morgan Kaufmann Publishers.

Ghani, R., & Yang, Q. (2017). Data cleaning: Problems and current approaches. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(3), 1-19.

Wang, R. Y., & Strong, D. M. (1996). A Taxonomy of Dirty Data. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, ACM, pp. 147-156.

"National Plan and Provider Enumeration System," Centers for Medicare & Medicaid Services, [Online]. Available: https://nppes.cms.hhs.gov. [Accessed: Feb. 5, 2023].