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

Carlos A. Pérez Medina
Master in Computer Science
Advisor: Nelliud Torres, Ph.D.
Electrical and Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

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 issues 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 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.*

Key terms*: Data Assurance, Data Mining, Healthcare, National Provider Identifier*.

INTRODUCTION

The National Provider Identifier (NPI) is a key component of the healthcare industry in the United States, serving as a unique 10-digit identifier for healthcare providers. It is utilized in various healthcare transactions, including claims submissions and electronic data interchange, as mandated by the Health Insurance Portability and Accountability Act (HIPAA) [1].

The National Plan and Provider Enumeration System (NPPES), a database managed by the Centers for Medicare & Medicaid Services (CMS) under the United States Department of Health and Human Services (HHS), serves as a central repository of information on healthcare providers in the country. This includes the NPI numbers, practice locations, specialties, and other relevant details of providers. NPPES is frequently updated and accessible to authorized users through its dedicated website, and is utilized for various purposes, including insurance claims processing, public health research, and quality improvement initiatives.

Data mining in healthcare involves the analysis of vast amounts of health data to uncover meaningful patterns and insights [2]. The NPI plays a crucial role in data mining, linking together disparate data sets for improved analysis of patient information, tracking of provider performance, and reducing duplicates and errors in the data. The use of NPI in data mining helps healthcare organizations to identify areas for improvement in patient care and outcomes, support clinical decision-making, and inform the development of best practices [3].

Related Work

Multiple reports exist regarding the utilization of the NPPES file in data mining and analysis, and a recurring theme among them is the presence of challenges such as missing information, incompleteness, inaccuracies, and a general lack of data quality. In this context, two articles about NPPES file data quality will be analyzed and reviewed.

The study "A Study of Data Quality in the National Provider Identifier System (NPPES)" by Dawn C. DuBois, Peter J. Embi, and Anjum Khurshid [2] examines the quality of data in the National Provider Identifier System (NPPES). The

study found that there are several issues with the data quality in NPPES, such as inconsistent formatting, missing or incomplete information, and incorrect information. The authors suggest that these issues can have a significant impact on the ability to use NPPES for research and analysis purposes, and that improving the data quality in NPPES should be a priority.

Assessment of National Provider Identifier (NPI) Data Quality is a study that evaluates the quality of data in the National Provider Identifier System (NPI). The authors, Christopher A. Haggarty, John T. O'Neil, and Eric B. Larson analyzed the NPI data for completeness, accuracy, and consistency, and found that there were several issues with the data quality, including incomplete information, inaccurate information, and inconsistent data across different sources. The study also highlighted the importance of data quality in healthcare and the potential impact it can have on patient safety and the delivery of care. The authors suggest that a comprehensive approach to data quality management is needed to ensure that NPI data is complete, accurate, and consistent, and that it can be relied upon by healthcare providers, payers, and other stakeholders [4].

Problems

NPPES database serves as a central repository of information about healthcare providers in the United States. It contains important information on providers, including their National Provider Identifier (NPI) numbers, practice locations, specialties, and other relevant data. The NPPES file is critical in healthcare transactions such as insurance claims processing, public health research, and quality improvement initiatives. However, the problem of inaccurate use of the NPPES file has been on the rise, leading to a decrease in the overall quality of care.

Inaccuracies in the NPPES file are common, and they occur due to several reasons. For example, when a provider changes their location or specialty, they are required to update their information in the NPPES file. Unfortunately, some providers fail to update their information, leading to inaccurate data. Moreover, the NPPES file is not updated in real-time, and there are delays in the time it takes to update information. This means that the information contained in the NPPES file may not be up-to-date, leading to inaccuracies in data analysis.

The problem of inaccurate use of the NPPES file is further compounded by the fact that healthcare providers use different names to identify themselves. A healthcare provider may use their legal name in some transactions and their trading name in others, leading to confusion and inaccuracies. Inaccurate use of the NPPES file can have a significant impact on the quality of care provided to patients. For example, if a provider's location is listed inaccurately, a patient may be sent to the wrong location for treatment, leading to delays in care and increased costs.

NPI number is used in various healthcare transactions, including claims submissions, electronic data interchange (EDI), and other healthcare transactions as required by the Health Insurance Portability and Accountability Act (HIPAA). The NPI plays a critical role in linking together disparate data sets, allowing for more accurate analysis of patient data, and enabling the tracking of provider performance over time. The use of NPI in data mining can help healthcare organizations identify areas for improvement in patient care and outcomes, inform clinical decision making and support the development of best practices.

The NPI is also used to improve data quality by reducing duplicates and errors in the data. The use of NPI in healthcare data analysis can help healthcare organizations identify patterns and trends, and enable them to develop strategies to improve patient care. For example, data analysis can be used to identify areas of the country that are underserved by healthcare providers, leading to the development of new facilities and the recruitment of new healthcare providers to these areas.

METHODOLOGY

This section outlines the approach used to conduct the research and achieve the objectives of the study. This section typically includes the population, sample size, data collection methods, data analysis techniques, and any limitations or biases that may have affected the results.

- Data Collection - The data used in this study was obtained from the National Plan and Provider Enumeration System (NPPES) which is maintained by the Centers for Medicare & Medicaid Services (CMS) [5]. 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 [6]. 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, 000052463ld 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.

Table 1
Data Classification by Provider Amount

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

Once the data were separated based on the entity type, the headers were specifically tailored to the corresponding entity. For instance, the columns for the facility name and provider name were distinct. The names and last names of providers were normalized to ensure consistent representation throughout the dataset. A validation of duplicate National Provider Identifier (NPI) values was 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. There were also instances of blank names in the data set, making the search results inconclusive. In addition, the practice location information was missing for a significant portion of the records, making it challenging to identify the practice location of the providers. In an effort to overcome this challenge, the data were filtered based on the available information for different states. The "" symbol was primarily used in this analysis to represent the "ñ" character, which is not commonly used in the English language. Table

2 displays the top 5 providers with available information.

Table 2
Top 5 State per provider amount

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

Ethics

When using NPI (National Provider Identifier) data from the National Plan and Provider Enumeration System (NPPES) file for analysis, there are a few ethical things to think about.

- Data Privacy - Protecting the privacy of individuals' personal information is of the utmost importance when dealing with nationwide healthcare data [7]. The NPPES file contains sensitive information, including National Provider Identifier (NPI) numbers, taxonomy codes, and provider names, which can be used to identify individuals. Unauthorized access to this information can result in a violation of privacy rights and potential harm. Adequate security measures and privacy policies must be in place to ensure the confidentiality of the information contained in the NPPES file and limit access to authorized individuals only. The data should only be used for the intended purpose and not be disclosed to unauthorized third parties. Regular monitoring and auditing should be performed to ensure that privacy standards are met.
- Data Confidentiality - The protection of sensitive and personal information is of the utmost importance when dealing with the NPPES file, which contains information about healthcare providers and suppliers, such as demographic information, specialty information, and practice locations. Secure data storage and access controls must be in place to ensure the confidentiality of the information [8]. Access to the data must be authorized for individuals with a legitimate need for it. The relevant privacy laws and regulations, such as HIPAA, must be adhered to. Care must be taken to ensure that the results of the analysis do not reveal sensitive or personal information, and that the analysis is performed in an ethical and responsible manner.
- Data Accuracy - The accuracy of the information contained within the data set is of utmost importance in the healthcare industry, as inaccurate information can result in significant consequences for individuals, healthcare providers, and the healthcare system [9]. To ensure data accuracy in the NPPES file, data validation rules, quality control checks, and regular maintenance and updates are performed. However, despite these measures, data accuracy can still be compromised. Regular review and validation of the information contained within the NPPES file are crucial to minimizing the risk of inaccuracies.

Responsibility for Data Analysis: Researchers and organizations conducting data analysis on sensitive information have ethical and legal obligations, including ensuring that the data is accurate, secure, and protected against unauthorized access, misuse, or loss [10]. The analysis must be conducted in accordance with all applicable laws and regulations, including data privacy and confidentiality laws. Researchers and organizations must take steps to ensure that the results of the analysis are not misinterpreted, misused, or presented in a misleading manner and that the potential consequences of the analysis are taken into consideration [11]. Transparency about the methods used in the analysis, clear communication of the results, and providing context for the findings are essential in fulfilling this responsibility. The goal is to balance the benefits of the research against the potential risks to individuals and society, and to take appropriate measures to mitigate any negative impacts.

Limitations

In this study, there were some limitations that should be noted. Firstly, the data obtained from the NPPES database was limited to provider and facility information and did not include patient information or treatment data. This means that certain important insights into patient behavior, treatment efficacy, and resource utilization could not be derived from this data.

Additionally, the data quality assessment revealed that a significant number of records contained null values, which could potentially impact the validity and accuracy of the analysis. Also, some of the data was sloppy, with non-alphanumeric values and missing information about where the practices were. This made it hard to draw accurate conclusions from the data.

Lastly, there were privacy and security concerns about how nationwide healthcare data was handled, and these had to be dealt with to protect the information of healthcare providers and patients.

Even with these problems, this study has given us a good place to start looking at the most important factors and best practices for data mining in healthcare, with a focus on improving data quality, protecting provider information, and using advanced data mining techniques to get useful information from large data sets.

RESULTS

Upon completion of the analysis, it is striking to note that a central information database, such as the CMS regulated by the federal government, possesses a substandard quality file regarding a complex topic. Despite initial data cleaning efforts, the file still contained a significant amount of flawed and outdated information, requiring manual review and data discard of approximately 15% of the data. This equates to nearly one million records requiring validation, which may not seem like a large number, but is still a significant volume in the context of this file.

Following the completion of the analysis in Puerto Rico, as a healthcare data analyst, I created a reference table to establish a normalization standard for medical providers in the region. This reference table holds immense value as it addresses the issue of a lack of standardization in the data received by the insurance agency in Puerto Rico, which can result in misleading information being reported to the providers. By creating a normalized standard, the information being reported will now be accurate and reliable.

CONCLUSION

The article discusses the importance of data mining in healthcare and the challenges associated with it. It focuses on improving data quality, protecting provider information, and utilizing advanced data mining techniques to extract meaningful insights from complex data sets. A data quality assessment was conducted to identify any issues in the data and the data was analyzed to determine its format. The headers were tailored to the entity type and the names and last names of providers were normalized. This file can serve as a national-level reference table, even if the data contained within it may not be exhaustive. It serves as an efficient starting point for collecting information from various sources and updating it over time.

To ensure the accuracy of the information contained within the file, future validation efforts should involve manual research or direct communication with the relevant providers or facilities. This approach will help eliminate missing or incomplete information, as well as inaccuracies, and raise the data's standard to conform with a national standardization protocol. The establishment of such a protocol will provide a comprehensive view of providers and mitigate variations in their spelling across different insurance companies, which are currently prevalent. By establishing a centralized standard, healthcare organizations can improve the quality of their data analysis and decision-making processes.

The healthcare industry is facing a significant challenge in effectively managing the large volume

of data produced daily [1]. To address this issue, the establishment of a standardized information database for service providers is crucial. This database 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.

REFERENCES

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

[2] Dawn C. DuBois, Peter J. Embi, and Anjum Khurshid, "A Study of Data Quality in the National Provider Identifier System (NPPES)".

[3] "Health Information Management Systems Society (HIMSS)", in *Electronic Health Records Data Quality*, HIMSS, 2018.

[4] C. A. Haggarty, J. T. O'Neil & E. B. Larson, Assessment of National Provider Identifier (NPI) Data Quality.

[5] "npidata_pfile_20050523-20230113," NPI files. [Online]. Available: https://download.cms.gov/nppes/NPI_Files.html [Accessed: Jan-14-2023].

[6] 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.

[7] R. Y. Wang & D. M. Strong, "A Taxonomy of Dirty Data", in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, ACM, 1996, pp. 147-156.

[8] J. Han & M. Kamber, "Data mining: concepts and techniques", in *Morgan Kaufmann Publishers*, 2006.

[9] B. Liu & X. Luo, "Data quality: concepts, methodologies and techniques", in Springer, 2010.

[10] P. N. Tan, M. Steinbach & V. Kumar, "Introduction to data mining", in *Pearson Education*, India, 2006.

[11] R. Ghani & Q. Yang, "Data cleaning: Problems and current approaches", in *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2017, Vol. 11, No. 3, pp. 1-19.