Improving Healthcare Provider Directories with Availity

Rishi Kumar (rk639), Matt Lim (sl2533), Helen Ma (hm385), Hana Mariappa (hm559),
Yuqing Wu (yw2465), Frank Yang (by93)

ORIE 4999 Final Summary Report

Professor David A. Goldberg

May 25, 2023

Abstract

This project focused on the importance of maintaining accurate healthcare provider directories.. Inaccuracies in such directories can lead to decreased patient satisfaction and potential fines from regulatory bodies. Existing research has highlighted Natural Language Processing (NLP) and data science techniques as potential solutions to improve accuracy, and this project proposes to build on this by implementing a scoring algorithm. Four datasets were used for analysis: NPPES, CMS, Healthgrades, and Zocdoc. Methods proposed included using Google Places API or locality sensitive hashing for string address matching. Based on the consistencies of the string address matching, we can then assign the scores to each address and rank our confidence level on the accuracy of the data. Initial manual comparison of data revealed several issues, including discrepancies between mailing and practice addresses, inconsistencies in names, and the handling of professionals with multiple practice locations. Note, we are unable to implement the whole steps in our proposed methods, i.e the step for scoring and weighting addresses. Future work will include implementing and testing the proposed scoring model and further analyzing the reasons behind inconsistencies.

Introduction

Healthcare provider directories serve to keep track of important information needed for patient care access, transactional exchange, insurance coverage, and other actions that require up-to-date records. Provider information can change quickly as people relocate or rebrand. A review done by Center for Medicare & Medicaid Services (CMS) on 54 Medicare Advantage Organizations (MAOs) found that about 45.1% provider directory locations listed were inaccurate [1]. If databases are not updated, patient satisfaction may decrease as their search for care becomes increasingly difficult, and CMS may fine organizations for using false information. Therefore, an accurate provider directory is important for increasing efficiency and reducing costs so clients can easily find providers.

About Availity

This semester we were tasked with a specific problem that concerned Availity. Availity is one of the biggest healthcare clearinghouses, offering multiple advanced solutions to different providers and vendors nationwide. Solutions include free portal connection to payers, submission of electronic transactions, revenue cycle optimization, and patient access management. The main task that our team focused on was the exchange of transactions between healthcare providers and patient payers. Through Availity, providers can simply access a platform and get connected to a network of real-time information exchange with payers. To do this, Availity must rely on up-to-date provider directories. Since Availity cannot change the databases, inconsistency in directories becomes a problem that needs to be addressed.

Background Reading

1. *Facilitating accurate health provider directories using natural language processing*

This paper [1] presents a natural language processing (NLP) technique that compares the National Plan and Provider Enumeration System (NPPES) and the Connecticut Department of Public Health. The NLP technique is able to identify inaccuracies correctly in many cases, but there are some limitations in which people's names are changed and failure to identify identity when there is a middle name included. The NLP also could not identify providers listed with a degree out of its existing scope. The NLP technique presented consists of a specialty taxonomy module and an entity recognition module that attempts to separate the providers based on their taxonomy, a scoring module that determines how accurate the information is, and an update module to generate accurate information with the combined sources.

2. *Data Science Techniques to Improve Accuracy of Provider Network Directory*

This paper [3] discusses various data science techniques to improve accuracy of the provider network directory. The provider network directory contains data about addresses, phone number, and specialities, and whether or not they are accepting new patients. The paper also mentions that 66% of healthcare inaccuracies were about the practice location. Traditionally, payers use the outreach method which includes phone calls, newsletters, and emails; however, this is very costly and inefficient. The proposed model was an intake scoring method. Intake

scoring is a Levenshtein-based distance calculation for address accuracy that computes scores for provider location and applies a decision tree that buckets the scores as high, medium, and low. This also utilizes a predictive model where they use labeled data from a call center data and predicts probability on whether the provider is practicing or not practicing at the stated location. This combines logistic regression, XGB boost, and random forest models. After six iterations of this model, they were able to create up to 79% accurate data.

3. *Natural Language Interface for Process Mining Queries in Healthcare*

This paper [2] refers to the idea that a business process management (BPM) can be used for healthcare data and provider management. This relates to the notion that workflow automation increases efficiency for daily business tasks. The healthcare provider databases contain information that is difficult to analyze using structured query language (SQL). Process mining can be used instead to extract information that is not easily obtained. The authors proposed a two part query engine, a process mining engine and a NLP pipeline. Process mining tools can monitor process patterns where it extracts information from event logs programmatically. However, the limitation is where it requires the user knowledge of APIs and components to interact with the process mining engine. The NLP pipeline makes data accessible for users without needing to program process mining queries. Then, users can enter queries into the GUI interface to obtain information. The model can extract a list of entities and query hints which can return to the user. This uses rule-based semantic parsing which includes the subtree pattern matching framework.

Datasets Used

Much of our time this semester consisted of finding datasets we wanted to verify the practice location of the provider across. We ultimately narrowed down to analyzing these four datasets: NPPES, CMS, Healthgrades, and Zocdoc. For the former two datasets, we were able to download csv files located on their websites, while for the latter two datasets, we used web scraping to obtain the data. An overview of each is given below.

1. *NPPES*

The NPPES database is maintained by the Centers for Medicare and Medicaid Services (CMS), which is a federal agency within the U.S. Department of Health and Human Services. The system assigns a unique National Provider Identifier (NPI) to each healthcare provider or organization.

2. *CMS*

CMS is the federal agency that runs the Medicare, Medicaid, and Children's Health Insurance Programs, and the federally facilitated Marketplace.

3. *Healthgrades*

Healthgrades is an online platform which offers a comprehensive database of healthcare professionals, facilities, and patient reviews to help individuals make informed decisions about their healthcare.

4. *Zocdoc*

Zocdoc is an online platform that allows users to find and schedule appointments with healthcare providers in their area.

Our Methods

Through the various readings we did, we concluded that implementing a scoring algorithm would be ideal. After web scraping and attaining the data from the datasets, we needed to clean up the data to remove the excess data and reformat to a standard format so it would be easy to compare against the datasets. We wanted to include identifiers (including name, phone number, and NPI number) and the provider location(s) only. After, we wanted to compare all six possible pairings of the datasets against each other. More specifically, for each provider name, we needed to check if their associated locations matched up across datasets. After finding a pattern of inconsistencies or consistencies, we could assign a corresponding "weight" to each data set. If a pair did not match, the weights would cancel out as a linear combination of both weights. If one pair matched more consistently, then the whole data set would be affected. We would go back and flag the providers which met the threshold score.

Some problems we recognized in this method were that multiple practice locations for one provider were often displayed inconsistently across the datasets. Another factor we considered was that because we were using online platforms, Healthgrades and Zocdoc, rural doctors may have less of a presence on these sites compared to urban doctors. Taking those factors into account, we wanted to assign weights to different regions geographically and providers that have recorded multiple practice locations in one dataset.

Unfortunately, due to the time limit, we are not able to complete everything we had in mind. We narrowed down the datasets and acquired the data, which we kept in a GitHub repository (see Appendix B). For address matching, we had two methods. We brainstormed a possibility of using Google Places API to compare the addresses. Since the same address from different datasets can have slightly different formats, we could not directly compare the address String. Thus, we used the Google Places API to get the geolocation, longitude, and latitude. If two addresses have similar geolocation within a certain threshold, say within 3 digits after the decimal, then the two addresses are considered to be the same address, and thus there is a consistency, and vice versa. This is shown in Figure 1, where we were able to test the model's ability to identify similarity of addresses. This approach was promising, and we have included the actual implementation in the GitHub repository. Using the machine learning and NPI models already implemented by Google was much more effective than creating our own ML models. Alternatively, the locality sensitive hashing (LSH), a way to match similar strings against each other, was also promising. The LSH uses a hash function to map similar strings close to each other and then find the neighbors of certain strings (in this case, addresses that we wanted to find). We have done some research on LSH, but we did not finish the actual implementation. We concluded that method 1 was more effective than method 2. However, we realized that method 1 could have limitations when two provider addresses are very close to each other. For example, if providers A and B are in the same building and next to each other, our method may not be able to

distinguish this as an unequal match. So, method 2 could be an alternative method during those exceptions. We were unable to get to the actual scoring or weighting process.

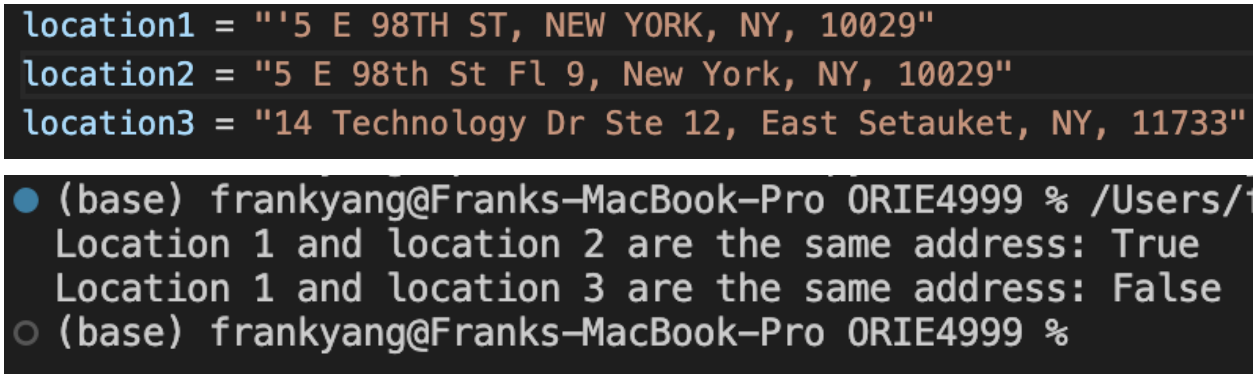


Figure 1: String matching result using Google Places API (code in Github). Rounded to thousandths place. This shows the Google Places API method to be very promising, though more testing with more data is needed.

Main Findings

Initially, before writing code to sort the various datasets, we attempted to match by hand. As there are six team members, we were each responsible to match ten healthcare professionals creating a dataset of 60 manually reviewed records. We each picked 10 names out of the NPPES data set at random, and then searched via their name on the NYS licensure website. A portion of this outcome is shown in Figure 2. Out of the 60, 12 failed to match across the NYS licensure website and the NPPES data set. We looked into why these discrepancies were occurring and revealed that many were mailing addresses versus practice addresses. There were also problems when there was an inconsistency with names (middle name, maiden name, etc.) Some professionals practiced at multiple practice locations which created confusion in the datasets and websites as sources only presented one location. Another factor contributing to limitations was that reliance on names did not always provide unique records. If a name was common, there were sometimes hundreds of entries associated. This pushed us to find another identifier to perform the matching. Additionally, there were some taxonomies of healthcare professionals that existed in some databases, but did not exist in another. For example, NPPES considered medical school students as healthcare professionals while the NYS licensure did not.

Last Name	First Name	Taxonomy Code	Corresponding Profession	License Number	Address	City	Comments
Kelly	Alexandra	235Z00000X	speech-language pathologist (058)		25 LYNBROOK CT	STATEN ISLAND	matched
Rulison	Casey	163W00000X	registered professional nurse (022)		22 EUCLID AVE	CORTLAND	matched

Campbell	Rohan	163W00000X	registered professional nurse (022)		9347 202ND ST	HOLLIS	matched
Graepel	Shannon	101YM0800X	licensed mental health counselor (018)		3815 VOORHIS LN	SEAFORD	name/license number did not come up on verification search
Huang	Meixian	171100000X	acupuncture (025)		107 NORTHERN BLVD STE 406	GREAT NECK	verification search address: GUILFORD CT
Geithner	Elise	363LP0200X	nurse practitioner - pediatrics (038)		555 W END AVE APT 4E	NEW YORK	verification search address: BURLINGTON VT
Yi	Ann	221700000X	creative arts therapist (005)		340 E 29TH ST APT 3I	NEW YORK	matched
Heineman	Christine	225700000X	massage therapist (027)		560 LANCER CT APT B2	DEPEW	matched
Carmiencke	Amber	225X00000X	occupational therapist (063)		2601 OCEANSIDE RD	OCEANSIDE	matched
Hughes	Kathleen	164W00000X	nurse, practical (010)		344 CHARLTON RD	BALLSTON SPA	matched

Figure 2: Sample of the results done by manual matching.

Conclusion

This project underscores the significance of maintaining accurate provider directories in healthcare, a process that is notably challenging due to providers' changing circumstances, inconsistencies in reporting, and differences in data collection methods across databases. Through our initial data analysis and exploration of methods such as the use of Google Places API or locality sensitive hashing for string matching, we have made strides towards improving data integrity in this crucial sector. Our preliminary manual matching of records highlighted critical areas for improvement, including differentiating between mailing and practice addresses, managing inconsistencies in provider names, and accounting for professionals with multiple practice locations. While we were unable to fully implement our proposed scoring model due to time constraints, the research and preliminary work conducted provide a promising foundation

for future exploration. This ongoing research will not only address the pressing need for accuracy in healthcare provider directories but also contribute to improved patient satisfaction, regulatory compliance, and overall healthcare service efficiency.

Going forward, we hope to implement the scoring model, conduct additional rounds of data matching, and perform a deeper analysis of the reasons behind discrepancies in provider location data. Additionally, after matching multiple rounds, we want to sort the records into buckets based on characteristics of the providers to identify the reason behind the occurrence of inconsistencies in practice locations. Whether that be geographic region, taxonomy of the healthcare professional, or even the name of the professional, we believe that further analysis of provider directories could suggest which factors are influential. For address matching, there is a possibility of using locality sensitive hashing to be an alternative to using Google Places API. An actual implementation of the LSH might be helpful to compare. Then, the next step would be to score and weighting the provider address. The computation formula to calculate the score and how to assign the weights will also require some further research.

References

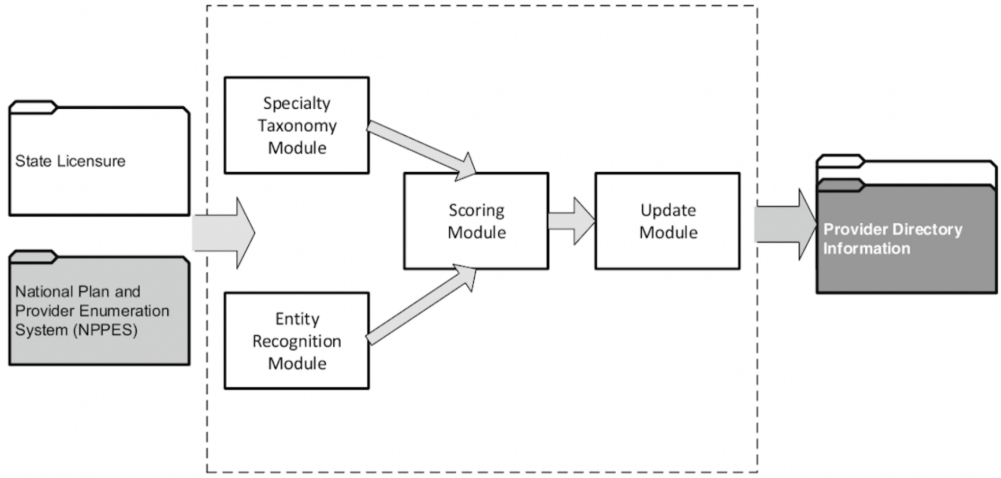
[1] Cook, M. J., Yao, L., & Wang, X. (2019). Facilitating accurate health provider directories using natural language processing. *BMC Medical Informatics and Decision Making*, *19*(S3). https://doi.org/10.1186/s12911-019-0788-x

[2] Yeo, H., Khorasani, E., Sheinin, V., Manotas, I., An Vo, N. P., Popescu, O., & Zerfos, P. (2022). Natural language interface for process mining queries in Healthcare. *2022 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata55660.2022.10020685

[3] Kandasamy, P., Raji, D., & Arun, S. (2019). Data Science techniques to improve accuracy of provider network directory. *2018 IEEE 25th International Conference on High Performance Computing Workshops (HiPCW)*. https://doi.org/10.1109/hipcw.2018.8634423

Appendix A

NLP model used in *Facilitating accurate health provider directories using natural language processing.*



Appendix B

GitHub repository containing all data processing codes, datasets mentioned, and string matching code.

https://github.com/whoisfrankyang/Cornell-Availity-Team.git