Exploring Data Quality and Variable Characteristics in Diabetes Readmission Analysis

Group 1:

Adeyinka Ajayi

Avinash Bunga

Ebenezer Amo (Captain)

Kwadwo Appiah

Master of Science in Information Systems and Business Analytics

PARK UNIVERSITY

CIS611HOS1P2024 Introduction to Business Analytics

Professor: Timur Rakhimov

February 4, 2024

Exploring Data Quality and Variable Characteristics in Diabetes Readmission Analysis

Introduction

In this project, our team delved into understanding diabetes readmissions using three key documents: the Diabetes Data Dictionary, Diabetes Readmission Exercise, and the Diabetes Profile. We examined the dataset's characteristics and quality through the Diabetes Profile tool. Our task was to analyze variable types, assess data quality, identify crucial data for understanding readmissions, and reflect on our findings and the analytical process. This paper presents our collaborative effort to explore and interpret the dataset, aiming to uncover insights that could lead to better management strategies for diabetes patients and reduce hospital readmissions. Through this concise analysis, we share our journey, findings, and the collective insights gained towards improving patient care outcomes.

Data Types

In data analysis, variables are the characteristics or attributes of the data that can vary or take on different values. Variables can be classified into different types based on the nature and level of the data they represent. In this text, we will explain the definitions and examples of three common types of variables: categorical, numeric, and Boolean.

Categorical Variables

Categorical variables sort information into specific groups or categories based on some qualitative or nominal property. For instance, in a dataset of hospital patients, some categorical variables could be:

- Race: This variable divides the patients into groups based on their race, such as Caucasian, African American, Asian, or Hispanic.
- Gender: This variable divides the patients into gender-based groups, such as male, female, or nonbinary.

- Age: This variable divides the patients into groups based on their age range, such as 0-10, 11-20, 21-30, and so on.

Categorical variables help us categorize patients to see how different groups might have different readmission rates or other outcomes of interest.

Numeric Variables

Numeric variables represent quantitative or numerical data that can be measured or counted. For example, in a dataset of hospital patients, some numeric variables could be:

- Time in hospital: This variable shows how many days a patient stayed in the hospital.
- Number of lab procedures: This variable tells us how many lab tests were done for a patient.
- Blood glucose level: This variable measures the amount of glucose (sugar) in a patient's blood.

Numeric variables are vital for measuring aspects of a patient's hospital stay and treatment, as well as their health status and risk factors.

Boolean Variables

Boolean variables are variables that have only two possible values: yes or no, true or false, 1 or 0. They are a special type of categorical variable that reflects the existence or absence of a certain condition or feature. For example, in a dataset of hospital patients, some Boolean variables could be:

- Diabetes medication: This variable tells us if a patient was prescribed diabetes medication or not.
- Change in medication: This variable indicates whether there was a change in any diabetic medication for the patient.
- Readmission: This variable shows whether the patient was readmitted to the hospital within a certain period.

Boolean variables help us quickly see which treatments were used and if any changes were made, making it easy to understand patient care patterns and outcomes.

Diagnostic Details: Present and Additional

The dataset includes specific diagnosis codes with diverse values. Diagnosis codes are standardized codes that identify and classify diseases and health problems. The dataset utilizes the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) system, which is a widely used coding system in the United States. Some examples of diagnosis codes and their meanings are:

- Diagnosis 1: Primary diagnosis, coded as the first three digits of the ICD-9-CM code, with 818 distinct values.
- Diagnosis 2: Secondary diagnosis, coded as the first three digits of the ICD-9-CM code, with 923 distinct values.
- Diagnosis 3: Additional secondary diagnosis, coded as the first three digits of the ICD-9-CM code, with 954 distinct values.

It is crucial for comprehending primary and secondary health issues that influence readmission risk to analyze the diagnosis codes and their frequencies in the dataset.

The dataset also includes some other diagnostic details that provide more information about the patient's health status and test results, such as the number of diagnoses, glucose serum test result, and A1c test result.

Additional Helpful Details That Would Be Beneficial

While the current dataset provides comprehensive diagnostic information, additional details would enhance the analysis and help identify more factors that affect readmission rates among diabetic patients. Some examples of additional details that could be useful are:

- Specific treatment procedures: This could include information about the type and frequency of surgical operations, medication dosages, and therapeutic interventions that the patient received during the hospital stay.
- Severity of diagnoses: This could include information about the severity or stage of each diagnosis, such as mild, moderate, severe, or critical.
- Timing of diagnoses: This could include information about when each diagnosis was made, such as before, during, or after the hospital stay.
- Interactions between diagnoses: This could include information about how different diagnoses might relate or contribute to each other, such as synergistic, antagonistic, or independent effects.
- Comorbidities or chronic conditions: This could include data about the patient's comorbidities or chronic conditions, such as hypertension, heart disease, kidney disease, depression, etc.
- These additional details would provide a more comprehensive understanding of the factors influencing readmission rates among diabetic patients.

Data Presence:

Which variables present in the Diabetes Profile file seem of particular importance in exploring the Diabetes Readmission question? Why?

When we look into why diabetic patients might have to come back to the hospital, certain pieces of information in the Diabetes Profile file stand out because they play a significant role:

1. **Time in Hospital**: This tells us how long a patient stayed in the hospital. A longer stay can mean more serious health issues, which might lead to returning to the hospital after discharge.

2. **Number of Lab Procedures and Diagnoses**: This shows how many tests and health problems a patient has. More tests and problems can mean a patient's health situation is complicated, increasing the chance of returning to the hospital.
3. **Demographic Factors (Age, Race, Gender)**: These details help us understand who the patients are. Different ages, races, and genders can affect how often patients return to the hospital for various reasons, including how they approach healthcare.
4. **Medication Variables (Diabetes Medication, Change of Medications)**: How diabetes is treated with medicine is crucial. If a patient's medication does not change much, it might mean their diabetes is under control. Frequent changes might show their health is unstable, affecting their risk of returning to the hospital.
5. **Readmitted**: This fact tells us if and when a patient had to return to the hospital. It helps us see patterns and figure out which patients might need more help to avoid coming back.
6. **Admission Type, Discharge Disposition, and Admission Source**: These pieces of information give us clues about a patient's journey in the healthcare system, like how quickly they needed care or where they went after leaving the hospital. This can help us understand why patients might return.
7. **Clinical Test Results (Glucose Serum, A1c)**: These tests measure how well diabetes is being managed. Bad results might mean a patient needs more help managing their diabetes to avoid coming back to the hospital.
8. **21 features for medications**: This detailed info about medicines helps us see if patients are sticking to their treatment plans or need changes to their medicines, which is critical to keeping them healthy and out of the hospital.
9. **Diagnoses (Diagnosis 1, Diagnosis 2, Diagnosis 3)**: Knowing the main and other health issues patients have helps us see what other problems they are dealing with,

which might make managing diabetes harder and increase their risk of returning to the hospital.

10. **Patient Number**: This unique ID helps us keep track of each patient's information over time, which is important for understanding their health journey.

Data Presence: What is not present that would be helpful or that one might expect to be present?

The dataset gives us a lot of useful information, but there are still some key pieces missing that could help us understand more about why diabetic patients end up back in the hospital. Adding these details could improve how we look after patients and work to keep them from needing to come back.

- **Post-Discharge Patient Behavior**: Knowing what patients do after they leave the hospital, like if they follow their diet and exercise plans, can give us important clues. This helps us see where patients might need more guidance or support to stay healthy.
- **Detailed Patient Medical History**: A more complete look at a patient's past health problems would make it easier to see if ongoing issues could lead to them being readmitted. This full history is important for understanding each patient's unique health challenges.
- **Post-Discharge Medication Adherence**: It is important to know if patients are taking their medications correctly after they go home. Not following medication instructions is a big reason people might have to return to the hospital. This info could help doctors and nurses determine which patients need extra help managing their medications.
- **Complete Weight Data**: The dataset is missing 97% of the weight data. Having this data for everyone could help us study how being overweight or underweight affects the chance of being readmitted. Since weight plays a significant role in managing

diabetes and overall health, understanding its effects could lead to better care for each patient.

- **Payer Code**: This is critical for understanding a patient's insurance coverage, but 52% of this data is missing. If we had complete and accurate details about everyone's insurance, we could better understand how a patient's financial situation and insurance coverage affect their access to medications and care after leaving the hospital. This knowledge could help doctors and nurses spot potential challenges patients might face in managing their diabetes and avoiding readmission.

Incorporating these data points, especially with the mentioned percentages of missing information, would offer a more holistic approach to analyzing readmission factors. It would also enable healthcare providers to address the broader needs of diabetic patients, aiming for improved outcomes and reduced hospital returns (Agrawal, 2018; Kumari, 2023).

Data Quality:

What is the general quality of this dataset?

The general quality of the dataset appears to be adequate, based on an examination of four objective data quality dimensions: completeness, accuracy, consistency, and relevance (Sebastian-Coleman, 2010). These dimensions reflect the extent to which the data is present, correct, uniform, and suitable for the intended analysis or research questions. Additionally, the dataset was compared with external sources or domain knowledge for validation. Below are the explanations of how this assessment was made:

How do you make that assessment?

Completeness: Completeness refers to the extent to which all required data is present (Sebastian-Coleman, 2010). High levels of completeness suggest that the dataset is more reliable and suitable for analysis. For the diabetics profile data, there are 4073 missing cells, which make up 0.1% of the data. This is a minute number, so the data quality is high.

Accuracy: Accuracy pertains to the correctness of the data (Sebastian-Coleman, 2010). It involves identifying and addressing any errors, outliers, or inconsistencies in the dataset. This can be achieved by cross-validating the data with external sources or through data cleaning and validation processes. An external source used was the Center of Disease Control and Prevention (CDC) data on diabetics available at CDC website. Comparing both data shows accuracy, especially in terms of gender and race.

Consistency: Consistency relates to the uniformity of data formats, units, and values within the dataset (Sebastian-Coleman, 2010). It involves ensuring that data is recorded and stored in a standardized manner, and that there are no conflicting or contradictory entries for the same entity. The diabetic dataset is consistent and uniform, as each variable follows a specific type/format.

Relevance: Relevance refers to the suitability of the data for the intended analysis or research questions (Sebastian-Coleman, 2010). It involves assessing whether the variables and records in the dataset align with the objectives of the analysis and whether any extraneous or redundant information is present. The Diabetes Data Dictionary and the Diabetes Readmission Exercise can provide insights into the relevance of each variable for exploring the diabetes readmission question.

The general quality of the diabetes dataset can be assessed by considering its completeness, accuracy, consistency, and relevance to the research questions at hand. This assessment may be further informed by established frameworks and indicators for evaluating the quality of diabetes care data (IMF, 2012; Data Ladder, 2022; Vitalflux, 2020).

Data Character

The Data Character section examines critical variables in the Diabetes Profile dataset, which contains information on diabetes surveillance data at national, state, and county levels and by age, sex, race/ethnicity, and education (CDC, 2022). We explore how these variables relate to readmissions among diabetic patients, which are costly and often associated with poor health outcomes (Alper et al., 2023). We also identify how these variables affect patient outcomes and healthcare utilization patterns, essential for developing predictive models and interventions to reduce readmission rates and improve patient care.

Time in Hospital:

Characteristics: This variable is a numerical value that shows how long the patient stayed in the hospital. It can range from 1 day to several weeks, depending on the patient's condition.

Importance: This variable indicates how acute the patient's illness is and how much medical care they need. Healthcare resource utilization is a primary measure that healthcare providers use, which may affect the likelihood of readmission. Previous studies have found that more extended hospital stays are associated with higher readmission rates for specific conditions and procedures (Jiang & Hensche, 2023).

The number of Lab Procedures and Diagnoses:

Characteristics: During the patient's hospitalization, healthcare professionals performed lab procedures and diagnoses, then counted using discrete numerical values. The medical reports demonstrate how much diagnostic evaluation the healthcare professionals conducted and how complex the patient's medical condition is.

Importance: More lab procedures and diagnoses may suggest a more complicated medical situation requiring more monitoring and management. This complexity may increase the risk of readmission and need further investigation. For example, one study found that patients with diabetes with more than 10 lab tests during their hospital stay had a higher readmission rate than those with fewer tests (Soh et al., 2020).

Medication Variables (Diabetes Medication, Change of Medications):

Characteristics: These variables are categorical, showing the type of diabetes medication prescribed and whether there was a change in medication during the hospital stay.

Importance: Medication management is crucial in diabetes care, affecting patient outcomes and readmission risks. Knowing the specific medications prescribed and any changes made helps to understand treatment effectiveness, adherence, and possible complications. Changes in medication may indicate adjustments to control the patient's condition, which could affect readmission rates. For instance, one study found that medication change upon admission was associated with lower readmission rates among diabetic patients (Mukhopadhyay et al., 2017).

These variables provide essential insights into the patient's medical condition, treatment regimen, and healthcare utilization, which are all relevant for understanding readmission risks in diabetic patients. By analyzing the characteristics of these variables, we set the stage for a comprehensive analysis of the dataset and the creation of effective interventions to improve patient outcomes.

Most Crucial Variable

The number of lab procedures and diagnoses is the most crucial variable for subsequent analysis based on the description of the data character section and the three variables listed. The reasons are as follows:

- Reflection of Complexity: This variable reflects the complexity of the patient's medical condition and the extent of diagnostic evaluation performed by healthcare professionals. These factors can significantly affect the patient's readmission risk and healthcare utilization patterns (Shang et al., 2021).

- Analytical Feasibility: The variable is discrete and numerical, which enables easy analysis and comparison with other variables. Descriptive statistics can summarize the data and reveal outliers or trends, such as mean, median, standard deviation, and frequency distribution. Moreover, inferential statistics, such as correlation, regression, and hypothesis testing, can clarify the relationship between this variable and others, such as readmission status, length of stay, medication change, and comorbidities (Soh et al., 2020).
- Supported by Research: Previous studies have shown the importance of this variable as a predictor of readmission risk among diabetic patients. For example, research indicates that patients with diabetes who undergo more than 10 lab tests during hospitalization have higher readmission rates than those with fewer tests (Zhang et al., 2019). Patients who undergo more lab procedures, have extended stays, and receive more diagnoses tend to have positive correlations between these factors (Zhao et al., 2019). These findings emphasize the clinical relevance of this variable and its potential usefulness in developing predictive models and interventions to reduce readmission rates and improve patient outcomes (Wu et al., 2014).

Therefore, it is advisable to prioritize analyzing the number of lab procedures and diagnoses. Furthermore, comparative assessments with the other variables, such as time in the hospital and medication variables, can provide insights into their interaction and mutual influence.

General Reflections

Examining Dataset Quality with the Diabetes Profile Tool

The Diabetes Profile tool is a self-administered questionnaire that assesses the social and psychological factors related to diabetes and its treatment (Elizabeth et al. Institute, n.d.). It helped us see how good our data was by showing us clearly where information was missing and if we had any repeated information. For example, it pointed out that much weight data was missing and helped us notice that some data stayed the same across different entries. This made it easier for us to spot where we might need to fix or fill in gaps in our data.

Understanding Data with the Diabetes Profile Tool

This tool provided information about the frequency of different health conditions and medication usage. It helped us better understand our data by highlighting patterns and crucial details we may want to investigate further. It was like a map of our data, guiding us to the information we needed. According to the American Heart Association (n.d.), understanding the data is the first step to preventing and managing diabetes.

Confidence in Descriptive Statistics for Analysis

We feel confident about using the statistics the Diabetes Profile tool gave us to start digging into our data. The tool alerted us about potentially complex data and highlighted areas for further exploration, providing a solid foundation for deeper analysis. Accu-Chek (n.d.) suggests that self-monitoring blood glucose levels is essential for diabetes management and analysis.

What to Do Next with Our Data Exploration

The insights we got from exploring our data are like clues on where to go next. We can use what we learned to improve our data, like fixing missing information. We can move on to more detailed studies, like looking at what might cause patients to return to the hospital. This first step of exploring our data helps us figure out the best ways to improve patient care and outcomes in the future. The American Diabetes Association (n.d.) offers a risk test and other resources to help people with diabetes take charge of their health.

Answer to Key Question from Stakeholders:

With our rich dataset, we can pinpoint critical factors leading to readmissions by analyzing variables such as the length of hospital stays, the number of lab procedures, and medication management. These insights allow us to identify high-risk patients and tailor interventions to manage diabetes more effectively, potentially reducing readmissions. Implementing targeted care plans based on these analyses can help address the specific needs of patients, improving their outcomes and minimizing the likelihood of returning to the hospital.

Conclusion

In our exploration of diabetes readmission data, we have dissected the Diabetes Data Dictionary, Diabetes Readmission Exercise, and the Diabetes Profile to deepen our understanding of what drives readmissions. This focused analysis revealed vital variables such as admission duration, lab procedures, and medication changes as significant predictors for readmission. We noted the dataset's general reliability but also recognized gaps that, if filled, could offer richer insights for managing diabetes more effectively.

Using the Diabetes Profile tool was pivotal in assessing data quality and understanding variable characteristics, boosting our confidence in using this data for exploratory analysis. Looking ahead, we suggest further research to address data gaps and a detailed study on how different health conditions interact, aiming to reduce readmission rates.

This project has sharpened our analytical skills and broadened our understanding, suggesting that leveraging detailed data analysis can significantly aid in crafting strategies to mitigate diabetes-related readmissions, ultimately enhancing patient care and reducing healthcare burdens.

References

Agrawal, S. (2018, July 20). *Why Is Data Analytics Important in Healthcare?* Dataversity. https://www.dataversity.net/data-analytics-important-healthcare/

Alper, E., O'Malley, T. A., & Greenwald, J. (2023). *Hospital discharge and readmission*. UpToDate. https://www.uptodate.com/contents/hospital-discharge-and-readmission

Accu-Chek. (n.d.). Self-monitoring of your diabetes. https://www.accu-chek.com.au/diabetes-basics/self-monitoring-your-diabetes

American Diabetes Association. (n.d.). Tools to know your risk. https://diabetes.org/tools-resources/tests-calculators

American Heart Association. (n.d.). Diabetes tools and resources. https://www.heart.org/en/health-topics/diabetes/diabetes-tools--resources

CDC. (2022). Diabetes data and statistics. https://www.cdc.gov/diabetes/data/index.html

Cleveland Clinic. (2021). Blood glucose test. Retrieved from https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test

Elizabeth Weiser Caswell Diabetes Institute. (n.d.). Survey instruments. https://diabetes.med.umich.edu/about/resources-health-professionals/survey-instruments

Healthline. (2020). Comorbidity: Definition, types, risk factors, treatment, and more. Retrieved from https://www.healthline.com/health/comorbidity

ICD9Data.com. (n.d.). The web's free ICD-9-CM & ICD-10-CM medical coding reference. Retrieved from http://www.icd9data.com/

Jiang, H. J., & Hensche, M. (2023). Characteristics of 30-day all-cause hospital readmissions, 2016-2020. HCUP Statistical Brief #304. https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.jsp

Kumari, S. (2023, September 8). *The Importance of Healthcare Data Analytics*. Aissel. https://www.aissel.com/blog/The-Importance-of-Healthcare-Data-Analytics

Markovič, R., Vladimir Grubelnik, Tadej Završnik, Helena Blažun Vošner, Kokol, P., Matjaž Perc, Marko Marhl, Matej Završnik, & Jernej Završnik. (2023). Profiling of patients with type 2 diabetes based on medication adherence data. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1209809

Mayo Clinic. (2018). A1C test. Retrieved from https://www.mayoclinic.org/tests-procedures/a1c-test/about/pac-20384643

Mukhopadhyay, A., Tai, B. C., See, K. C., Ng, W. Y., Lim, T. K., & Li, J. (2017). Risk factors and outcomes of very early readmissions in patients with diabetes. Journal of Diabetes and its Complications, 31(2), 346-351. https://doi.org/10.1016/j.jdiacomp.2016.11.006

Rothe, U., Manuwald, U., Kugler, J., & Schulze, J. (2020). *Quality criteria/key components for high quality of diabetes management to avoid diabetes-related complications*. Journal of Public Health. https://doi.org/10.1007/s10389-020-01227-w

Soh, J. G. S., Wong, W. P., Mukhopadhyay, A., Quek, S. C., & Tai, B. C. (2020). *Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis*. BMJ Open Diabetes Research & Care, 8(1), e001227. https://doi.org/10.1136/bmjdrc-2020-001227

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). *The 30-day hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers*. BMC Medical Informatics and Decision Making, 21, 57. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01423-y

Wu, S., Duan, R., & Wisdom, K. (2014). *Taking steps in the hospital to prevent diabetes-related readmissions.* American Nurse Today, 9(4), 18–23. https://www.myamericannurse.com/taking-steps-in-the-hospital-to-prevent-diabetes-related-readmissions/

Zhang, Z., Jiang, K., Shang, Y., Wang, L., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2019). *Implementation of machine learning algorithms to create diabetic readmission risk prediction models.* BMC Medical Informatics and Decision Making*,* 19, 240. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x

Zhao, Y., Li, Y., & Zhao, Y. (2019). *Predicting 30-day hospital readmissions for patients with diabetes*. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (pp. 1–6). IEEE. https://storm.cis.fordham.edu/~yzhao/tp/Publications/C14_HIMS_2019_readmission.pdf