

Project Assignment 1

Group 3: Lukas Blazsovsky, Vincent Rauchegger, Simon Wolffhardt

Contents

Dataset Description	2
Raw Dataset Variables	2
Data Quality Issues	4
Initial Analysis Questions	4
Exploratory Analysis	5
Within each primary diagnosis group, what are the top independent predictors of a readmission?	7
What are the key differences between early (<30 days) and late (>30 days) readmission?	9
How does a change in medication strategy change the readmission risk?	10
Summary	12

Dataset Description

The dataset describes diabetes patient admissions/encounters in several US hospitals over a period of 10 years. It consists of over 100,000 rows and 50 variables. Each record is made up of patient information like gender, age group and weight group but also variables describing the encounter like the length of stay, the admission source and the readmittance duration.

The dataset can be accessed via <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>.

Raw Dataset Variables

The initial dataset consists of the following columns (variables):

i Info

Some variables are not listed here, because their function is unknown or they are not relevant to our analysis.

Variable	Description	Possible Values
encounter_id	Every encounter with a patient has a unique identifying number.	Integers.
patient_nbr	Every patient has a unique identifying number.	Integers.
race	The race of the patient.	'Caucasian', 'AfricanAmerican', '?', 'Other', 'Asian', 'Hispanic'
gender	The gender of the patient.	'Female', 'Male', 'Unknown/ Invalid'
age	The age group of the patient.	'[0-10)', '[10-20)', '[20-30)', '[30-40)', '[40-50)', '[50-60)', '[60-70)', '[70-80)', '[80-90)', '[90-100)'
weight	The weight group of the patient.	'?', '[0-25)', '[25-50)', '[50-75)', '[75-100)', '[100-125)', '[125-150)', '[150-175)', '[175-200)', '>200'
admission_type_id	An ID mapping to the type of admission.	Emergency (1), Urgent (2), Elective (3), Newborn (4), Not Available (5), NULL (6), Trauma Center (7), Not Mapped (8),
discharge_disposition_id	An ID mapping to the discharge disposition.	Integer from 1-29. Examples are Discharged to home (1), Expired (11), Hospice / home (13), NULL (18)

Variable	Description	Possible Values
admission_source_id	An ID mapping to the source of admission.	Integer from 1-26. Examples are Physician Referral (1), Clinic Referral (2), Transfer from a hospital (4), NULL (17)
time_in_hospital	Number of days the patient stayed at the hospital.	Integers (number of days).
medical_specialty	What type of medical specialty was dealing with the patient.	'Pediatrics-Endocrinology', '?', 'InternalMedicine', 'Family/GeneralPractice', 'Cardiology', 'Surgery-General', 'Orthopedics', ...
num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, number_diagnoses	Several metrics counting certain actions performed by the hospital staff or prior patient history.	Integers.
diag_1, diag_2, diag_3	Diagnosis as ICD-9 code. (See https://en.wikipedia.org/wiki/List_of_ICD-9_codes)	ICD-9 code.
max_glu_serum, A1Cresult	Result of laboratory tests, if performed	e.g.: 'Normal', '>200', '>300', 'None'
metformin, repaglinide, nateglinide, chlorpropamide, ..., insulin,	Several medication statuses.	'No', 'Steady', 'Down', 'Up'
change	If the medication changed in any way.	'Ch', 'No'
diabetesMed	If the patient takes diabetes medication.	'Yes', 'No'
readmitted	If and after how many days a patient has been readmitted.	'NO', '<30', '>30'

Data Quality Issues

1. **Missing values:** Many records have missing values in certain fields, often, but not always, indicated by a question mark (?).
2. **Inconsistent naming of boolean categories:** Some variables have values that are binary but have named values, like 'Yes' or 'No'. The column 'change' has 'Ch' instead of 'Yes'.
3. **Categorical values that are represented by an ID:** Some values are represented by an ID and need to be corresponded to their string representation via an additional provided CSV-file.
4. **Categorical values with overlap:** Columns like 'admission_source_id' have multiple values associated with incomplete or missing data like 'NULL', 'Not Mapped', 'Not Available' and 'Unknown/Invalid'.
5. **Limited number of detailed Diagnosis:** The dataset only includes up to three diagnosis per record even though there can be many more per encounter.
6. **Numerical values that have been categorized into value ranges:** Values like age or weight were anonymized by replacing them with the respective range, making these columns harder to process.

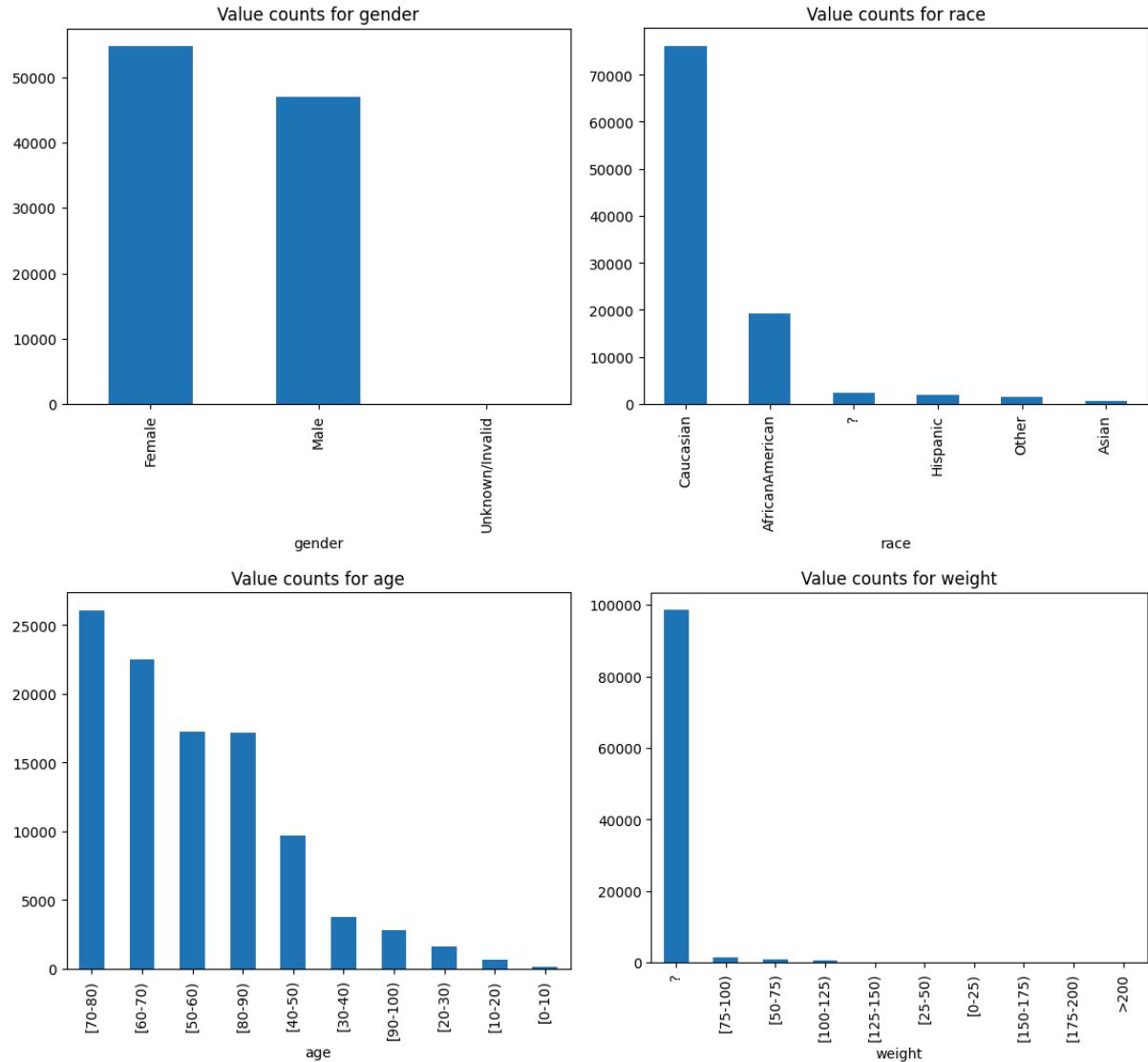
Initial Analysis Questions

These are the initial research questions we formulated:

1. What are the highest risk factors for patient readmission?
2. Which race, gender, or age-related differences are there?
3. What are the key differences between early and late readmission?
4. How does a change in medication strategy change the readmission risk?

Exploratory Analysis

First we want to get an overview of the dataset. We start by analyzing different patient attributes for each encounter.

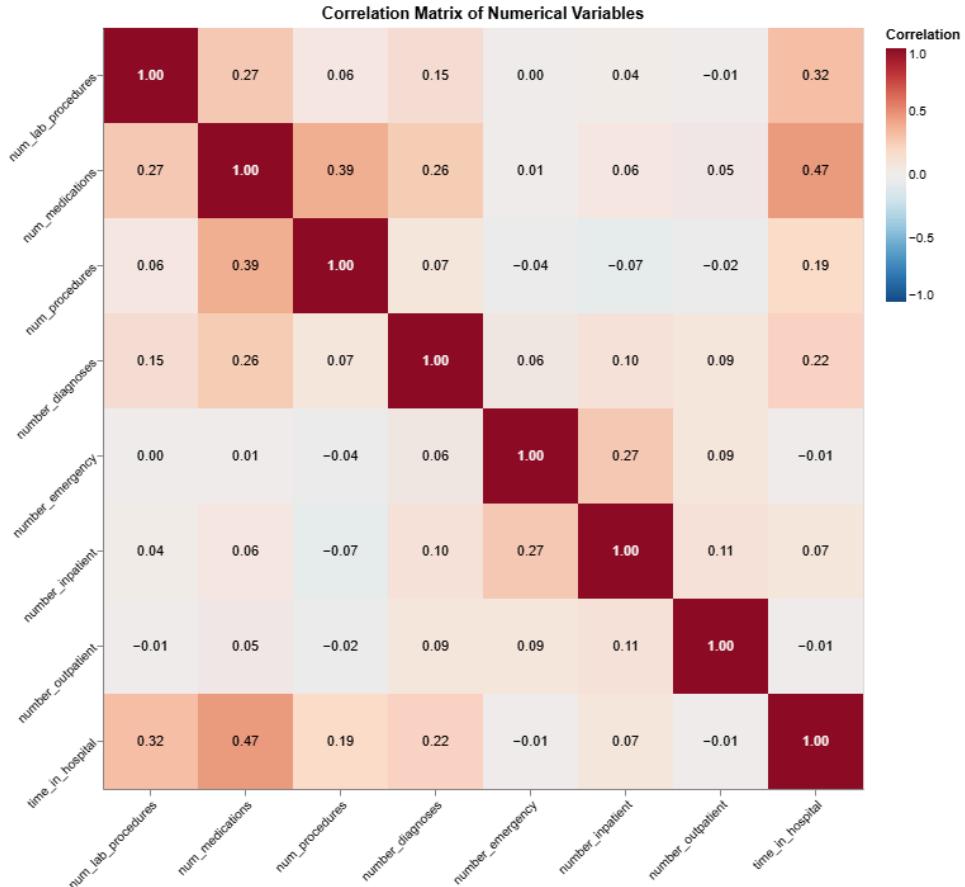


Let us analyze each attribute:

- **Gender:** The majority of encounters feature patients classified as female in the dataset.
- **Race:** Most encounters feature patients classified as Caucasian, followed by African American. This makes sense given the origin of the data. A significant portion of the dataset has missing values for this attribute.
- **Age:** The age distribution shows that most encounters feature patients between 50 and 80 years old.
- **Weight:** The weight attribute has a large number of missing values, which is probably because patients were not weighed consistently during their visits. Among the available data, most encounters feature patients in the weight categories between 75 and 125.

To test the reliability of the dataset we want to test some hypotheses regarding the variables. If we analyse the values of the variable `time_in_hospital`, `num_procedures` and `num_medications` we should see some correlation between them. The more procedures and medications a patient has, the longer they should stay in the hospital.

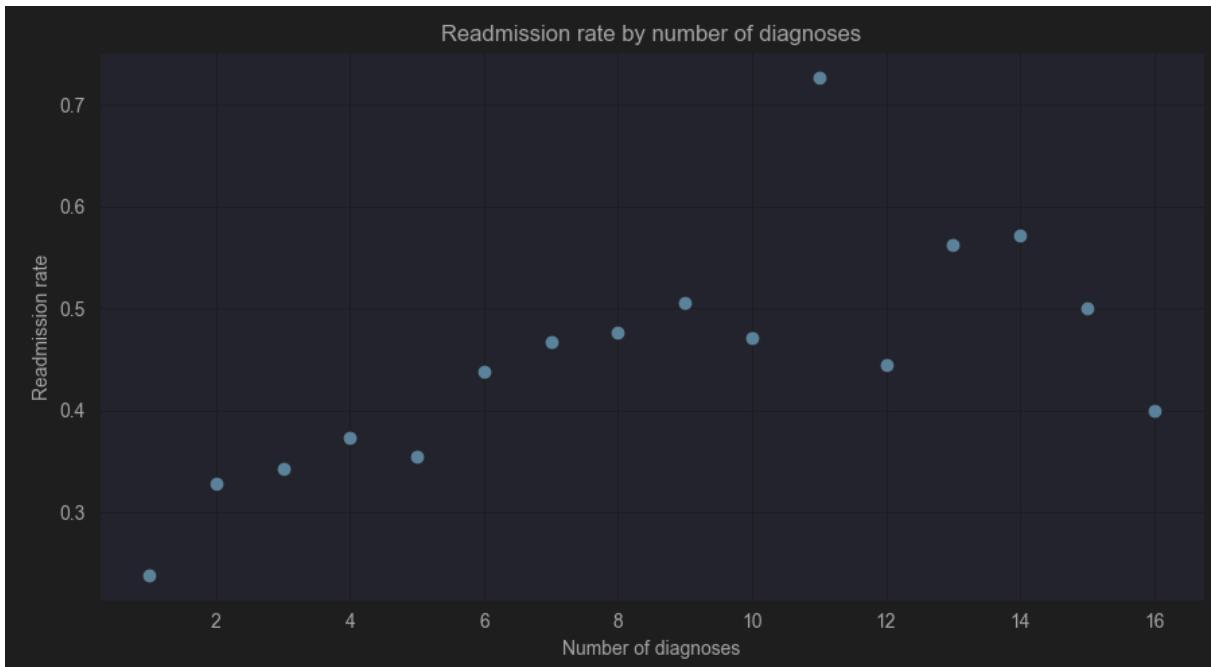
We check for this by exploring the numerical correlations in the dataset. The following correlation matrix shows the correlations between the values in the dataset (id values excluded).



This visualization shows that there are indeed correlations between the number of procedures/medications and the time spent in the hospital. Other correlations that can be seen are for example between the number of emergency and inpatient visits.

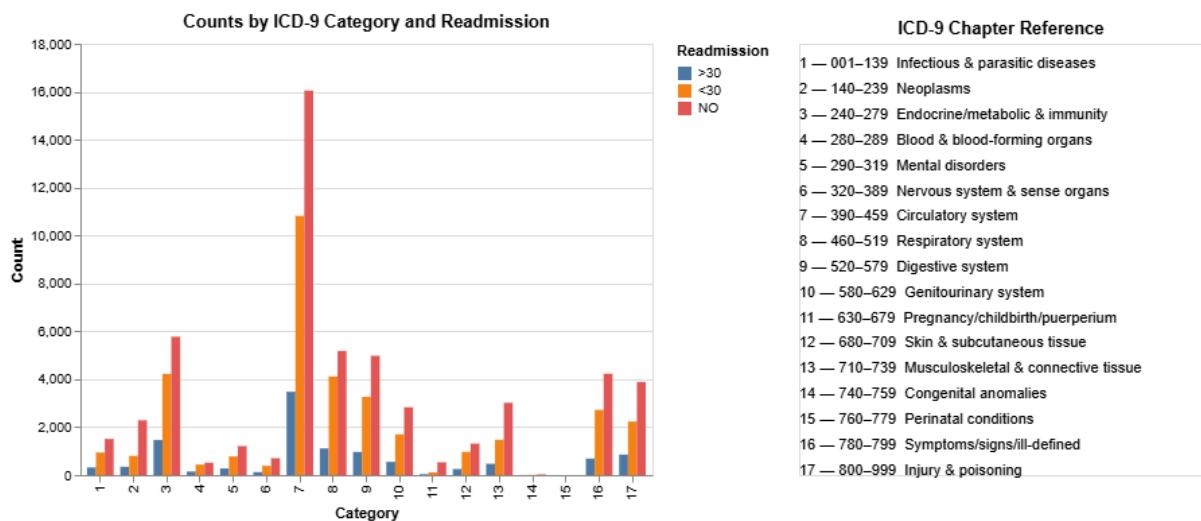
Within each primary diagnosis group, what are the top independent predictors of a readmission?

This scatterplot shows the readmission rate grouped by the number of diagnosis of a patient.



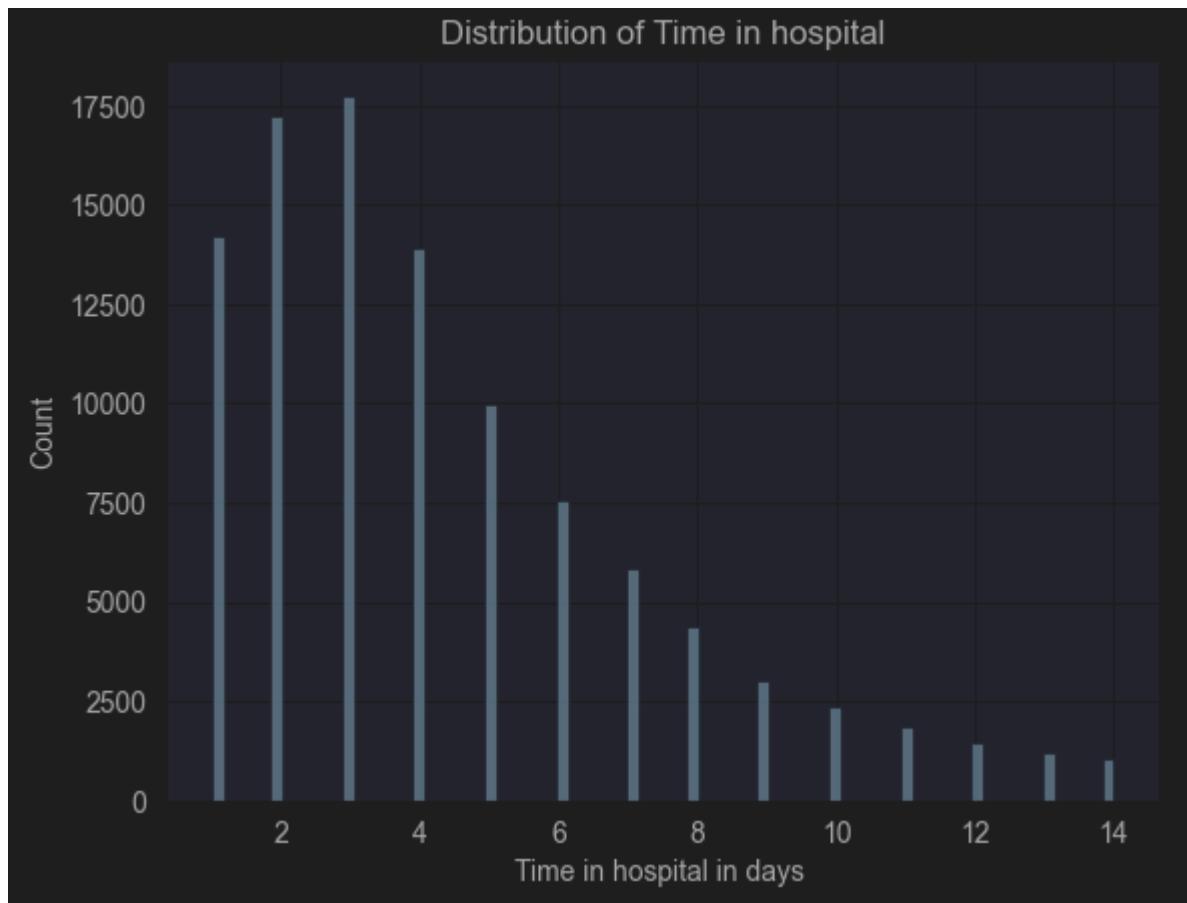
This plot shows a clear trend in readmissions as the odds of readmission increase directly proportional with the number of diagnosis. Now it would be interesting if there are some diagnosis types which increase the readmission odds more than others.

The following graph shows exactly this.



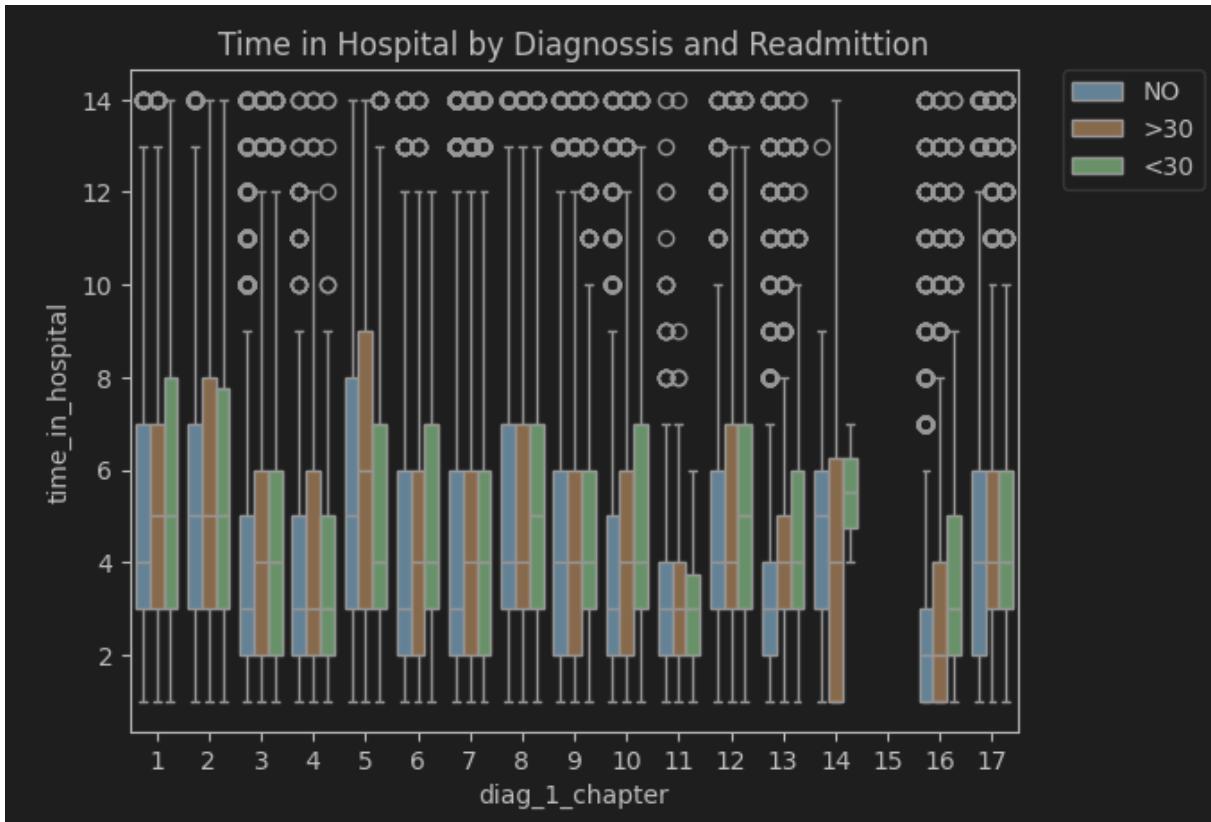
The main insight we gain here is that in all ICD-9 chapters the amount of people who are not readmitted is the highest. We can also see that the distribution is similar in all areas. Therefore this does not show us a clear trend

This Graph shows how many people spent how much time in the hospital.



A broad overview over how long patients are in the hospital is displayed here. The goal of this is to understand the more detailed stay duration visualization in the next graph better.

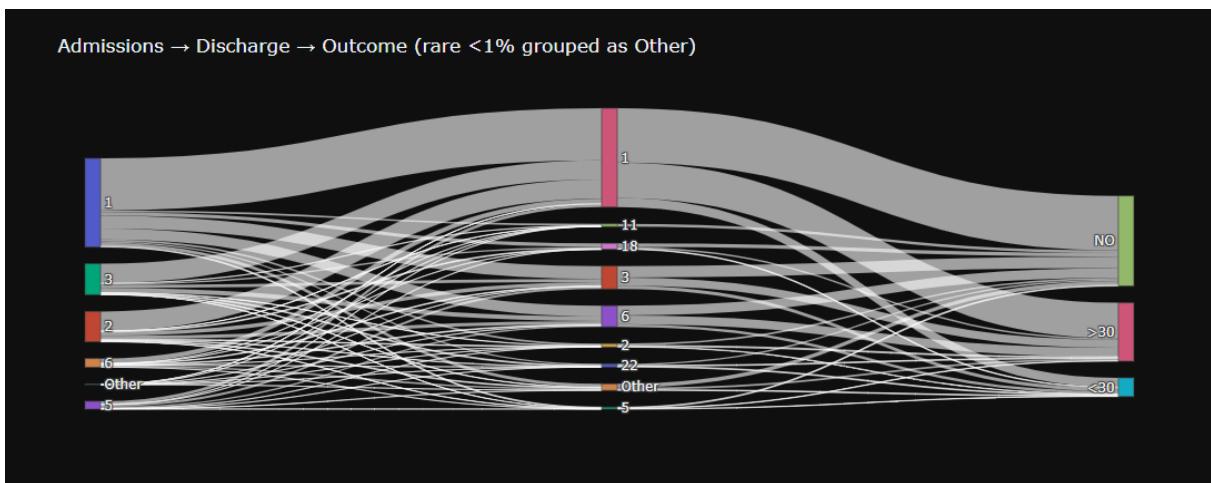
The next graph has the aim to show the correlation of diagnosis and the time spent in the hospital to readmissions



One can see that the time stayed in the hospital generally has a high variance with some outliers. In the chapter 16, Symptoms/signs/ill-defined, the time in hospital is very low for patients who are not readmitted but get visibly higher the sooner the readmission is. This can be investigated.

What are the key differences between early (<30 days) and late (>30 days) readmission?

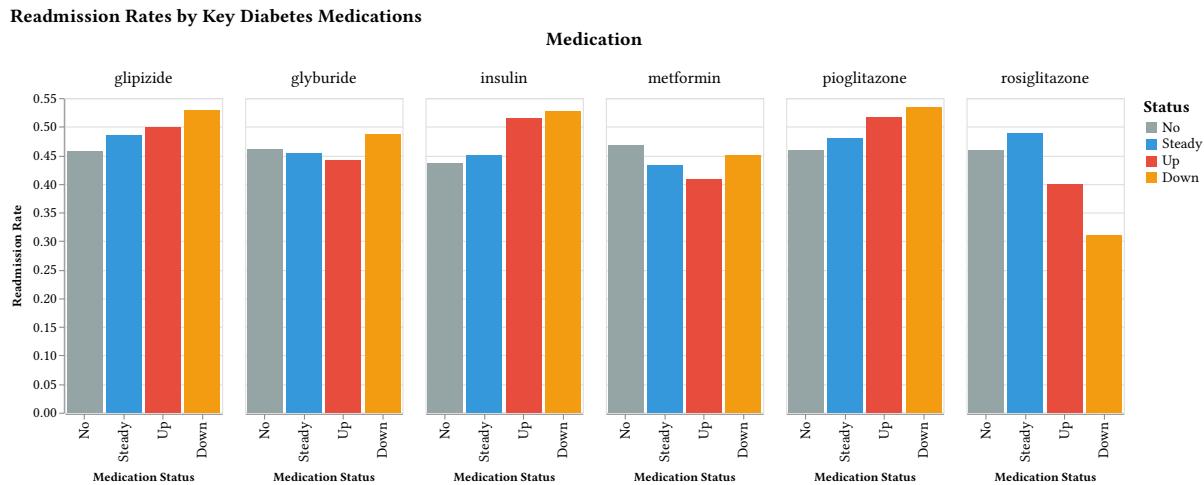
This Visualization shows a Sankey diagram where the patient flow from the first admission to the discharge type to the readmission is shown. Only patient flows over 1% are shown. The orders are grouped in others



We learn from this where patients most likely are admitted from and we can get an overview. Most patients are admitted through the emergency room or are admitted by another physician. But there is no clear trend visible that shows which kind of discharge type is most likely to result in readmission, as all look quite similar.

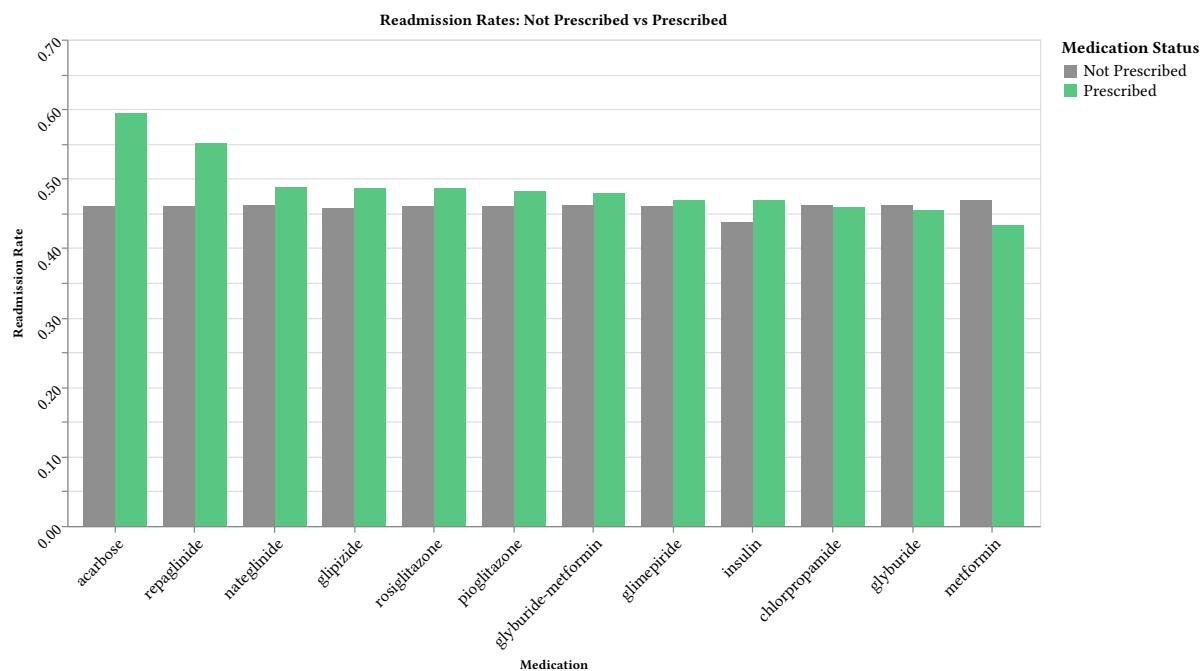
How does a change in medication strategy change the readmission risk?

Next we want to look at the effect of medication changes on readmission risk. In the following chart we can see the readmission rates by medication and change status.



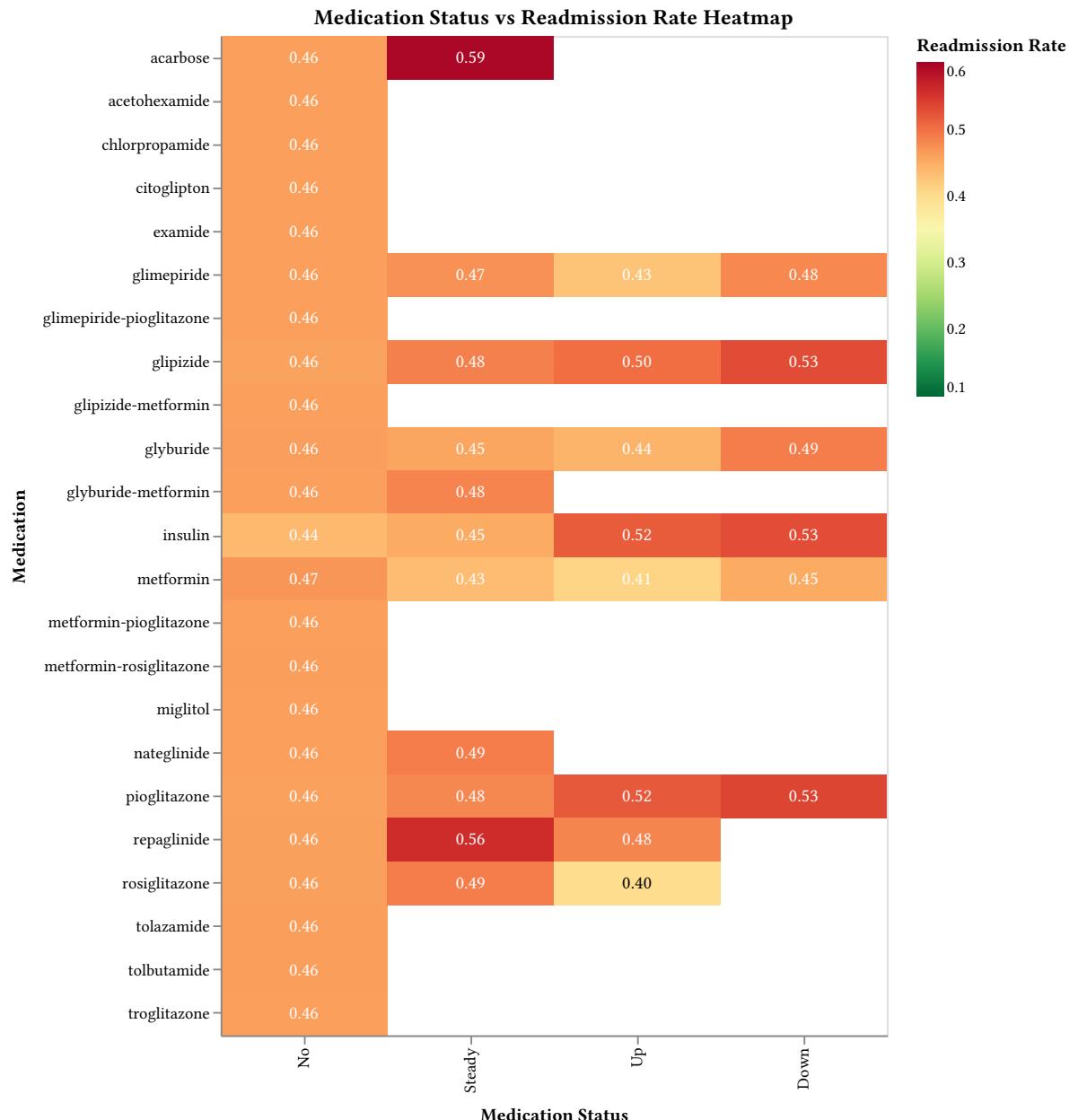
What is immediately apparent is that different medications have very different readmission rates depending on whether the medication was increased, decreased or stayed the same. For example, for Metformin, patients whose medication was increased ('Up') have a significantly lower readmission rate compared to those whose medication stayed the same ('Steady') or decreased ('Down'). This suggests that increasing Metformin dosage may be beneficial in reducing readmission rates.

For the next analysis we grouped the medication statuses into two categories: 'Prescribed' (which includes both 'Steady' and 'Up') and 'Not Prescribed'. This allows us to compare the overall effect of being on a medication versus not being on it at all.



From this chart, we can see that for most medications, being prescribed the medication is associated with a higher readmission rate compared to not being prescribed it. However, not being prescribed metformin, for example, is associated with a notably higher readmission rate compared to being prescribed it.

This heatmap shows the readmission rates for each medication based on whether it was never prescribed, decreased, increased, or stayed the same. It excludes status changes with fewer than 100 samples to ensure statistical significance.



Interestingly, increasing the dosage of rosiglitazone ('Up') is associated with a lower readmission rate compared to other medications. This would suggest that increasing rosiglitazone dosage may be particularly effective in reducing readmission rates. Notably, rosiglitazone is known to have cardiovascular side effects¹, which led to it being withdrawn from the market in several countries (e.g. the EU²).

⚠️ Warning

These findings should be interpreted with caution. Other factors may influence readmission rates and medication changes beyond the scope of this analysis.

¹<https://www.pharmawiki.ch/wiki/index.php?wiki=Rosiglitazon>

²<https://www.ema.europa.eu/en/medicines/human/EPAR/avandia>

Summary