

# 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 .....	4
Summary .....	4

## 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):

### 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 <a href="https://en.wikipedia.org/wiki/List_of_ICD-9_codes">https://en.wikipedia.org/wiki/List_of_ICD-9_codes</a> )	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. What are the key differences between early and late readmission?

## Exploratory Analysis

## Summary