Motivation

For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled?*

Hospitals have struggled with the management and treatment of patients with hyperglycemia who are admitted in non-intensive care units (ICU).

At the time of the creation of this dataset, there were "few assessments of diabetes care in hospitalized patients to serve as a baseline" for understanding diabetic patient care in non-ICU situations.

Who created the dataset (e.g., which team) and on behalf of which entity?

The dataset was created by Strack et al. (2014): a team of researchers from a variety of disciplines, ranging from computer science to public health, from three institutions (Virginia Commonwealth University, University of Cordoba, and Polish Academy of Sciences).

Composition

What do the instances that comprise the dataset represent?

Each instance in this dataset represents a hospital admission for diabetic patient (diabetes was entered as a possible diagnosis for the patient) whose hospital stay lasted between one to fourteen days.

Is any information missing from individual instances?

The features `Payer Code` and `Medical Specialty` have 40,255 and 49,947 missing values, respectively. For `Payer Code`, these missing values are reflected in the category *Unknown*. For `Medical Specialty`, these missing values are reflecting in the category *Missing*.

For our demographic features, we are missing the `Gender` information for three patients in the dataset. These three records were dropped from our final dataset. Regarding `Race`, the 2,271 missing values were recoded into the `Unknown` race category.

Does the dataset identify any subpopulations (e.g., by age, gender)?

Patients are identified by gender, age group, and race.

For gender, patients are identified as Male, Female, or Unknown. There were only three instances where the patient gender is *Unknown*, so these records were removed from our dataset.

Gender	Count	Percentage
Male	47055	46.2%
Female	54708	53.7%

For age group, patients are binned into three age buckets: *30 years or younger*, *30-60 years*, *Older than 60 years*.

Age Group	Count	Percentage
30 years or younger	2509	2.4%
30-60 years	30716	30.2%
Older than 60 years	68538	67.4%

For race, patients are identified as *AfricanAmerican*, *Caucasian*, and *Other*. For individuals whose race information was not collected during hospital admission, their race is listed as *Unknown*.

Race	Count	Percentage
Caucasian	76099	74.8%
AfricanAmerican	19210	18.9%
Other	4183	4.1%
Unknown	2271	2.2%

Preprocessing

Was any preprocessing/cleaning/labeling of the data done?

For the `race` feature, the categories of *Asian* and *Hispanic* and *Other* were merged into the *Other* category. The `age` feature was bucketed into 30-year intervals (*30 years and below*, *30 to 60 years*, and *Over 60 years*). The `discharge_disposition_id` was binarized into a boolean outcome on whether an patient was discharged to home.

The full preprocessing code is provided in the file `preprocess.py` of the tutorial GitHub repository.

Uses

Has the dataset been used for any tasks already?

This dataset has been used by Strack et al. (2014) to model the relationship between patient readmission and HbA1c measurement during admission, based on primary medical diagnosis.

The dataset is publicly available through the UCI Machine Learning Repository and, as of May 2021, has received over 350,000 views.