Data Intake Report

Name: Healthcare – Persistency of a Drug
Report date: July 25, 2022
Internship Batch: LISUM 10
Version:<1.0>
Data intake by: Jeeyeon Shon
Data intake reviewer:
Data storage location:
https://github.com/shonjeeyeon/DG_Week_8/blob/main/Healthcare_dataset.csv
(Original xslx file:
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view)

Tabular data details:
Healthcare_dataset.csv

Total number of observations	3,424
Total number of files	1
Total number of features	69
Base format of the file	.csv
Size of the data	892 KB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Ptid can be used to identify and remove duplicate observations
- The dataset has been deidentified already
- No missing values, however there are practical missing values such as 'unknown'. These values should be imputed appropriately (Will be discussed in the Deliverable Week 8 document)
- Most of the features are categorical; will need encoding to enable ML