Data Intake Report

Name: **HealthCare Dataset**
Report date: 05/19/2024
Internship Batch: LISUM32
Version: 1.0
Data intake by: Apurwa Bhausaheb Sontakke
Data intake reviewer:
Data storage location:
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view

Tabular data details:

Total number of observations	3,425
Total number of files	2 (sheets in the Excel file)
Total number of features	72
Base format of the file	Excel
Size of the data	898 KB

Proposed Approach

- Deduplication Validation (Identification):
  - Identify and eliminate duplicate records using unique patient IDs.
  - Maintain data integrity by cross-checking multiple identifiers (e.g., patient ID, date of birth).

- Assumptions:
  - All entries with the same patient ID pertain to the same individual.
  - Missing values can be addressed using statistical methods (mean/median) or model-based techniques.
  - Data quality issues like outliers can be managed through standard preprocessing methods.