Data Intake Report

Name: Persistency of a drug

Report date: 19/04/2025

Internship Batch:LISUM43

Version:<1.0>

Data intake by: MOHAMMADISTIYAK SHAIKH

Data intake reviewer: Data Glaciers

Data storage location: <https://github.com/iamistiyak/PersistancyOfDrugsByDG>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 3424 |
| **Total number of files** | 1 |
| **Total number of features** | 69 |
| **Base format of the file** | .xlsx |
| **Size of the data** | 900kb |

**Proposed Approach**

**1. Deduplication Validation (Identification):**

To identify and validate potential duplicate records in the dataset, the following multi-step approach will be used:

* **a. Define Uniqueness Criteria:**  
  Use the Patient ID as the primary key. Any records sharing the same Patient ID will be flagged for further inspection.
* **b. Row-Level Comparison (Non-ID Duplicates):**  
  Check for full or partial duplicate rows *excluding* the Patient ID, especially if multiple records exist for a single patient.
* **c. Near-Duplicates (Fuzzy Match):**  
  Use similarity checks (e.g., Levenshtein distance or cosine similarity) across high-cardinality fields such as:
  + Ntm\_Speciality, Race, Age\_Bucket, Region, etc.
  + Clinical features where overlap may indicate data re-entry
* **d. Frequency & Temporal Analysis:**  
  If timestamps or sequence IDs are available (not shown yet), sort and analyze the ordering to identify back-to-back entries that may represent duplicate intake or therapy episodes.
* **e. Manual Flagging Rules:**  
  Create rules like: if Gender, Race, Ethnicity, Region, and all comorbidity/risk flags are identical across patients, flag as potential duplicate for manual review.

**2. Assumptions for Data Quality Analysis:**

* **a. Single Record per Patient Assumed:**  
  Unless otherwise noted, we assume each Patient ID corresponds to a unique record, and there are no multiple episodes per patient in the data.
* **b. ‘Unknown’ Values Are Legitimate Categories:**  
  Entries marked as 'Unknown' (e.g., Change\_T\_Score) are treated as valid categorical levels unless further guidance suggests they are missing values.
* **c. No Temporal Leakage Expected:**  
  Features used are assumed to be either prior to or concurrent with therapy start (rxdate). We assume no data from after therapy conclusion leaks into predictive features.
* **d. Data Encoding Is Consistent:**  
  It is assumed that binary flags (e.g., Y/N) are used consistently across all columns. Any deviation (e.g., lower/uppercase or missing) will be standardized during preprocessing.