# **RxNorm Dataset Documentation**

## **Overview of RxNorm Files and Fields**

RxNorm is a standardized nomenclature for clinical drugs created by the National Library of Medicine (NLM). The dataset consists of several files, with the primary ones being RXNCONSO, RXNSTY, and RXNSAT.

## **1. RXNCONSO (Concept Names and Sources)**

Contains all concepts in RxNorm from various sources.

### **Key Fields:**

* **RXCUI** : Unique identifier for each concept
* **LAT** : Language of the term (ENG for English)
* **TS** : Term status
* **HIM** : Lexical unique identifier
* **STT** : String type
* **SUI** : String unique identifier
* **ISPREF** : Flag indicating if the term is preferred (Y/N)
* **RXAUI** : Unique identifier for each atom (term-concept pair)
* **SAUI** : Source asserted unique identifier
* **SCUI** : Source concept unique identifier
* **SDUI** : Source descriptor unique identifier
* **SAB** : Source abbreviation
* **TTY** : Term type
* **CODE** : Source-specific concept identifier
* **STR** : String (the actual term)
* **SRL** : Source restriction level
* **SUPPRESS** : Suppression flag (Y if obsolete, N if active)
* **CVF** : Content view flag

## **2. RXNSTY (Semantic Types)**

Contains semantic type information for each concept.

### **Key Fields:**

* **RXCUI** : RxNorm concept identifier
* **TUI** : Type unique identifier
* **STN** : Semantic type tree number
* **STY** : Semantic type name
* **ATUI** : Attribute unique identifier
* **CVF** : Content view flag

## **3. RXNSAT (Attributes)**

Contains attributes for concepts.

### **Key Fields:**

* **RXCUI** : RxNorm concept identifier
* **HIM** : Lexical unique identifier
* **SUI** : String unique identifier
* **RXAUI**: RxNorm atom identifier
* **STYPE**: Source type
* **CODE**: Source terminology concept code
* **ATN**: Attribute name
* **ATV**: Attribute value
* **SUPPRESS** : Suppression flag
* **CVF** : Content view flag

## **Term Types (TTY) in RxNorm**

### **Clinical Drug Terms**

* **SCD** (Semantic Clinical Drug): Contains ingredient + strength + dose form *Example: "Acetaminophen 500 MG Oral Tablet"*
* **GPCK** (Generic Pack): Grouped generic clinical drugs *Example: "Amoxicillin 500 MG Oral Capsule 30 Count Pack"*
* **BPCK** (Brand Name Pack): Grouped branded clinical drugs *Example: "Tylenol 500 MG Oral Tablet 50 Count Pack"*

### **Branded Drug Terms**

* **SBD** (Semantic Branded Drug): Brand name + ingredient + strength + dose form *Example: "Tylenol 500 MG Oral Tablet"*
* **BN** (Brand Name): Just the brand name *Example: "Tylenol"*
* **SBDC** (Semantic Branded Drug Component): Brand name + ingredient + strength *Example: "Tylenol 500 MG"*
* **SBDF** (Semantic Branded Drug Form): Brand name + ingredient + dose form *Example: "Tylenol Oral Tablet"*

### **Ingredient Terms**

* **IN** (Ingredient): The active ingredient *Example: "Acetaminophen"*
* **MIN** (Multiple Ingredients): Combination of ingredients *Example: "Acetaminophen / Codeine"*
* **PIN** (Precise Ingredient): Specific form of ingredient *Example: "Acetaminophen Sodium"*

### **Drug Components**

* **SCDF** (Semantic Clinical Drug Form): Ingredient + dose form *Example: "Acetaminophen Oral Tablet"*
* **SCDC** (Semantic Clinical Drug Component): Ingredient + strength *Example: "Acetaminophen 500 MG"*
* **DF** (Dose Form): Just the dose form *Example: "Oral Tablet"*

### **Others**

* **SCDG** (Semantic Clinical Drug Group): Clinical drugs grouped by ingredient and route *Example: "Acetaminophen Oral Product"*
* **DFG** (Dose Form Group): Grouped dose forms *Example: "Oral Product"*
* **SBDG** (Semantic Branded Drug Group): Branded drugs grouped by ingredient and route *Example: "Tylenol Oral Product"*

## **Source Abbreviations (SAB) in RxNorm**

* **RXNORM** : Concepts created by RxNorm itself
* **MTHSPL** : FDA Structured Product Labels
* **MMX** : Micromedex
* **VANDF** : Veterans Affairs National Drug File
* **MDDB** : Medi-Span Master Drug Database
* **NDFRT** : National Drug File - Reference Terminology
* **ATC** : Anatomical Therapeutic Chemical Classification System
* **MTHFDA** : FDA National Drug Code Directory
* **MSH** : Medical Subject Headings
* **SNOMEDCT\_US** : SNOMED CT US Edition
* **GS** : Gold Standard Drug Database
* **NDDF** : First DataBank National Drug Data File
* **MMSL** : Multum MediSource Lexicon
* **MTHCMSFRF** : CMS Formulary Reference File
* **MTHSPL\_2022\_02\_01** : FDA SPL specific version
* **USPMG** : USP Medicare Model Guidelines
* **FDB** : First Databank
* **DRUGBANK** : DrugBank database

## **Important Attributes (ATN) in RXNSAT**

* **AMBIGUITY\_FLAG** : Indicates if the term is ambiguous
* **RXN\_AVAILABLE\_STRENGTH** : Available strength of the drug
* **RXN\_QUANTITY** : Quantity information
* **RXN\_QUALITATIVE\_DISTINCTION** : Qualitative distinctions
* **RXN\_HUMAN\_DRUG** : Indicator for human use drugs
* **RXN\_ACTIVATED** : Activation date
* **RXN\_OBSOLETED** : Obsolete date
* **RXN\_STRENGTH** : Strength of the drug
* **NDC** : National Drug Code
* **SPL\_SET\_ID** : Structured Product Label set identifier
* **RXN\_IN\_EXPRESSED\_FLAG** : Flag for expressed forms of ingredients
* **RXN\_BOSS\_FROM** : Based-on relationship source
* **RXN\_BOSS\_TO** : Based-on relationship target
* **RXN\_VET\_DRUG** : Indicator for veterinary drugs

## **Semantic Types (STY) Common in RxNorm**

* **Clinical Drug**: Pharmaceutical preparations
* **Pharmacologic Substance**: Substances used for treatment
* **Organic Chemical**: Chemical compounds based on carbon
* **Biomedical or Dental Material**: Materials used in medicine or dentistry
* **Chemical Viewed Structurally**: Chemicals from structural perspective

## **Relationship Between Files**

1. **RXNCONSO** contains all concept names/terms
2. **RXNSTY** links concepts to semantic types using RXCUI
3. **RXNSAT** provides additional attributes for concepts using RXCUI and RXAUI

## **Common Data Challenges**

1. **Duplication** : Many-to-many relationships between concepts and attributes
2. **Ambiguity** : Some terms are ambiguous (check AMBIGUITY\_FLAG)
3. **Versioning** : RxNorm is updated monthly; concepts may be deprecated
4. **Size** : Complete joining of all files can result in billions of records

## **Approach for Data Processing**

1. **Filter First** : Start with specific TTYs and SABs based on your use case
2. **Process in Chunks** : Use memory-efficient processing for large files
3. **Selective Joins** : Only include necessary attributes
4. **Denormalize Carefully** : Convert attributes to columns instead of rows
5. **Check SUPPRESS Flag** : Filter out suppressed concepts (SUPPRESS='Y')
6. **Handle Ambiguity** : Pay attention to the AMBIGUITY\_FLAG attribute