An exploratory data science approach to understanding the OMOP data for use as phenotype in genotype-phenotype association studies

Xinlu Shi
nxw782@alumni.ku.dk
https://github.com/HauserGroup/Project_XS_PerMed

January 27, 2025

1 Introduction

1.1 Background

The integration of pharmacological, clinical, and genomic data holds immense promise in advancing personalized medicine. The primary aim of personalized medicine is to tailor drug prescriptions to individual patients, optimizing therapeutic efficacy while minimizing adverse effects. The UK Biobank offers a unique opportunity to explore this goal, leveraging its extensive dataset of over 55 million prescription records and whole-genome sequencing (WGS) data from 500,000 individuals. Such a dataset provides the foundation for uncovering complex relationships between drug utilization patterns, adherence, and genetic factors.

A core challenge in understanding drug utilization lies in the dynamic nature of drug prescriptions, particularly the patterns of drug switching and adherence. Drug switching—defined as the transition from one drug to another within or across therapeutic classes—serves as a proxy for clinical decision-making. Such decisions often reflect considerations of treatment efficacy, side effects, or evolving patient conditions. While adverse reactions are not explicitly encoded in most registries, drug-switching patterns offer indirect insights into these events and lay the groundwork for future genome-wide association studies (GWAS) on drug trajectories and adherence behaviors.

The OMOP Common Data Model provides a standardized framework for analyzing drug eras, which represent continuous or near-continuous periods of drug use. By employing this model, we can systematically identify drug switching patterns, quantify adherence gaps, and investigate the underlying factors

that drive these behaviors. This approach enables a more detailed understanding of individual treatment trajectories and their stratification by demographic and genetic factors.

1.2 Related Work

Recent advancements in data science methodologies and large-scale datasets have laid a strong foundation for understanding drug utilization patterns. Disease trajectory approaches, as detailed by Siggaard et al. (2020) in the Disease Trajectory Browser, showcase the power of population-wide data in uncovering longitudinal disease progression and multimorbidity patterns[5]. This methodology employs statistically significant directional diagnosis pairs to construct disease trajectories, emphasizing its utility in identifying common multimorbidity networks across diverse populations. The tool's ability to merge linear trajectories into disease trajectory networks makes it particularly relevant for stratifying patient subgroups and exploring complex patterns of disease progression.

Similarly, Haue et al. (2022) expanded on disease trajectory analysis by focusing on ischemic heart disease (IHD) patients. They identified length-two and length-three temporal disease trajectories, revealing how the timing and sequence of diagnoses delineate unique subpopulations within IHD cohorts. Their work provides a template for using temporal data to uncover insights into drug switching behaviors in pharmacological datasets[2].

Genomic predictors have also been a focal point in recent research. Kiiskinen et al. (2023) leveraged longitudinal medication purchase records to explore genetic associations with drug adherence and switching patterns in cardiometabolic diseases. Notably, they identified 333 independent loci associated with medication use behaviors, including polygenic risk scores (PRS) that predict switching and adherence patterns. This study highlights the potential of integrating genomic data with real-world prescription records to personalize therapeutic strategies[3].

The integration of whole-genome sequencing (WGS) data with prescription records, as showcased by Halldorsson et al. (2022), exemplifies the potential for high-resolution genomic datasets in studying drug utilization. By analyzing over 150,000 genomes from the UK Biobank, the study uncovered rare and common variants influencing drug response, paving the way for refined models of pharmacogenomics and personalized medicine[1].

These studies collectively reflect the evolving landscape of drug utilization research, emphasizing innovative methodologies like disease trajectory mapping, genome-wide association studies (GWAS), and the development of polygenic predictors. Their findings directly inform the current study's aim to explore drug switching patterns and adherence behaviors within the UK Biobank, bridging critical gaps between pharmacological and genomic research.

1.3 Objectives

The primary objectives of this study are threefold:

1. To delineate patterns of drug switching and adherence within the UK Biobank dataset, stratifying by demographic and genetic factors.
2. To evaluate the utility of pointwise mutual information (PMI) scores in identifying clinically meaningful drug switches.
3. To investigate how drug adherence, as measured by gap days, varies across therapeutic classes and its implications for patient outcomes.

By addressing these objectives, this work aims to establish a foundational framework for integrating pharmacological and genomic data, ultimately contributing to the broader goal of personalized medicine. The insights generated will also inform future studies, such as GWAS, aimed at understanding the genetic basis of drug utilization and adherence.

2 Method

2.1 Drug Era

The concept of a drug era is central to analyzing longitudinal drug utilization patterns and serves as a standardized representation of continuous or near-continuous drug use. In this section, we introduce the drug era framework, its components, and its relevance to the study of medication adherence and switching behaviors.

2.1.1 Definition and Key Components

A drug era is defined as a period during which a patient is exposed to a specific drug, allowing for short interruptions in treatment. The primary components of a drug era, as implemented in the OMOP Common Data Model [4], include:

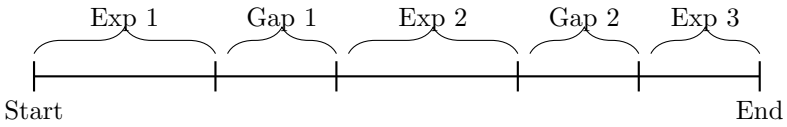
- **Person**: The individual associated with the drug era, identified by a unique person ID.
- **Drug Concept**: The specific drug or active ingredient being tracked, identified by a standardized concept ID.
- **Start Date**: The date on which the drug era begins, typically corresponding to the first recorded instance of drug use.
- **End Date**: The date on which the drug era ends, either due to discontinuation or a break in treatment exceeding the predefined gap threshold.
- **Gap Days**: The total number of days not covered by drug exposure records within the drug era, constrained by a threshold (30 days).

- **Drug Exposure Count**: The total number of distinct drug exposures within the drug era, reflecting adherence and re-initiations.

A single drug era aggregates multiple individual drug exposures when the gaps between them are below the predefined threshold. For instance, if a patient pauses drug use for less than 30 days before resuming the same medication, the exposure events are consolidated into a single drug era.

2.1.2 Example of Drug Era Construction

To illustrate the concept, consider the following example:



In this example, the patient begins with Drug Exposure 1, marking the start of Drug Era 1. During this era, a gap (Gap 1) occurs but does not exceed the predefined threshold, allowing the drug era to continue. Subsequent drug exposures, such as Drug Exposures 2 and 3, follow a similar pattern where the gaps (e.g., Gap 2) remain below the threshold. The total gap days for this drug era would be the sum of Gap 1 and Gap 2. However, if a gap were to exceed the threshold, a new drug era would be initiated.

2.1.3 Relevance to Adherence and Switching Analysis

Drug eras provide a comprehensive framework for analyzing medication adherence and switching patterns. By consolidating individual drug exposures, this approach minimizes noise in the data and captures meaningful patterns of drug utilization. Metrics such as gap rate, defined as the ratio of gap days to the total duration of the drug era, and switching patterns, which identify transitions between drug eras for different medications, can be calculated using this framework. This foundational structure is critical for the subsequent analysis of drug switching behaviors and adherence trends within large datasets, such as the UK Biobank.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) refers to the process of summarizing and visualizing key characteristics of the dataset to identify patterns, relationships, and potential issues. This section outlines the EDA conducted on the dataset to understand drug information and individual-level drug utilization.

2.2.1 Drug Information

The dataset includes drug-related identifiers such as ATC codes, ChEBI codes, and DrugBank IDs. ATC codes classify drugs based on their therapeutic use and chemical properties. ChEBI codes provide information on chemical entities of biological interest, while DrugBank IDs link drugs to detailed biochemical and pharmacological data. The intersections of these identifiers were studied to ensure consistency and integration across different coding systems.

2.2.2 ATC Code Component

Drugs were mapped to their corresponding ATC codes. For drugs associated with multiple ATC codes, all codes were retained to preserve comprehensive mapping. This approach ensures that all therapeutic categories relevant to a drug are included in the analysis, providing a complete representation of drug utilization patterns.

2.2.3 Individual-Level Drug Era

At the individual level, the analysis focused on drug utilization patterns. Key metrics include the total number of times each drug was taken, the number of individuals taking each drug, the number of times each person took each drug, and the total number of drugs taken by each person. These metrics provide insights into drug adherence, frequency of use, and polypharmacy across the population.

2.3 Drug Switch

2.3.1 Definition and Terminologies

After the conclusion of a drug era, a new drug era may begin, which constitutes a drug switch. Formally, a subsequent drug era is defined as drug era B following drug era A if both belong to the same individual and the start date of B occurs after the end date of A. Among all subsequent drug eras of A, the closest subsequent drug era is the one with the earliest start date.

A drug switch is denoted as (drug era A $\rightarrow$ drug era B), where drug era B is the closest subsequent drug era of A. The switch interval is defined as:

$$|\text{drug\_era\_start\_date}(B) - \text{drug\_era\_end\_date}(A)|.$$

For a drug switch (drug era A $\rightarrow$ drug era B), the corresponding drug concepts (drug concept of A $\rightarrow$ drug concept of B) define the drug switch pattern. Here, the drug concept of drug era A is called the *source drug*, while the drug concept of drug era B is referred to as the *destination drug*.

In the dataset, each drug era is evaluated to identify its closest subsequent drug era. If no subsequent drug era exists, this occurs either because the individual has only one drug era or all other drug eras for the same individual precede the end of this drug era. Note that this does not imply that the drug

era is the last chronologically. Furthermore, a single drug era may have multiple closest subsequent drug eras.

2.3.2 PMI Score on Drug Switches

Pointwise Mutual Information (PMI) is a statistical measure that quantifies the association between two events, capturing how much the probability of one event deviates from independence with the other. Mathematically, PMI between two events x and y is defined as:

$$PMI(x, y) = \log_2 \left[\frac{\Pr(x, y)}{\Pr(x) \cdot \Pr(y)} \right],$$

where $\Pr(x, y)$ is the joint probability of x and y, and $\Pr(x)$ and $\Pr(y)$ are the individual probabilities of x and y, respectively. PMI values indicate the strength of association: higher values suggest stronger relationships, while values close to zero imply independence. In the context of drug switches, PMI helps evaluate how meaningful a specific drug switch pattern is and how related two different drug switch patterns are.

PMI for a drug switch pattern (drug A $\rightarrow$ drug B) is defined as:

$$PMI(\text{A} \rightarrow \text{B}) = \log_2 \left[\frac{\Pr(\text{A} \rightarrow \text{B}) + \epsilon}{\Pr(\text{A} = \text{source drug}) \cdot \Pr(\text{B} = \text{destination drug}) + \epsilon} \right].$$

In this equation, $\Pr(\text{A} \rightarrow \text{B})$ represents the probability of observing the switch pattern (A $\rightarrow$ B), calculated as the number of switches with the pattern A $\rightarrow$ B divided by the total number of all drug switches. Similarly, $\Pr(\text{A} = \text{source drug})$ represents the probability that drug A is the source drug in any drug switch, calculated as the number of drug switches where drug A is the source drug divided by the total number of all drug switches. The term $\Pr(\text{B} = \text{destination drug})$ is calculated as the number of drug switches where drug B is the destination drug divided by the total number of all drug switches. To avoid issues such as division by zero or logarithms of zero, a small constant ϵ (set to 10^{-30}) is added to both the numerator and denominator.

Before conducting the analysis, drug switch patterns observed in fewer than 30 individuals were filtered out of the dataset. This preprocessing step ensures that the analysis focuses on statistically robust patterns and mitigates the bias of PMI toward infrequent events.

PMI values were then computed for each drug switch pattern to assess their significance. This analysis illustrates patterns with strong associations between specific source and destination drugs, providing insights into meaningful therapeutic transitions. By identifying these strong associations, the PMI metric aids in understanding the relationships between drugs and their switching behaviors.

To further explore the relationships between drug switch patterns, PMI was extended to measure the relatedness of two different drug switch patterns. The PMI for two switch patterns, switch 1 and switch 2, is defined as:

$$PMI(\text{switch 1, switch 2}) = \log_2 \left[\frac{\Pr(\text{switch 1, switch 2}) + \epsilon}{\Pr(\text{switch 1}) \cdot \Pr(\text{switch 2}) + \epsilon} \right].$$

Here, Pr(switch 1, switch 2) is the probability of a person having experienced both switch patterns, calculated as the number of individuals with both switch patterns divided by the total number of individuals. The terms Pr(switch 1) and Pr(switch 2) are calculated as the number of individuals who experienced the respective switch pattern divided by the total number of individuals. Again, ϵ is added to ensure numerical stability.

Given the large scale of the dataset and the study's focus, the analysis was restricted to drug switches within the ATC category for the nervous system (ATC code: N). This targeted approach provides insights into associations and patterns specifically relevant to nervous system drugs, which are of particular interest in this study.

2.3.3 Clustering on Drug Switch Patterns Based on PMI

To group related drug switches, the Louvain community detection algorithm was used. An undirected graph was constructed, where nodes represent drug switch patterns and edges connect pairs with PMI scores exceeding 4.0. Edge weights were set to the PMI values. This threshold reduced noise by focusing only on strongly related pairs, as low or negative PMI scores indicate weaker or exclusive relationships.

The Louvain algorithm optimized modularity, a measure of connection density within communities compared to random chance. By iteratively merging nodes, it identified clusters of drug switches with stronger internal connections than external ones. Each drug switch was assigned to a single community. Communities were evaluated by size, average internal PMI, and coverage (proportion of node pairs with significant PMI). These metrics assessed the strength and coherence of each cluster.

This method identified clusters of frequently co-occurring drug switches, revealing patterns linked to common therapeutic strategies or treatment pathways.

2.4 Gap Days

Gap days refer to the number of days not covered by drug exposure records that make up a drug era. This metric is a useful indicator of a patient's adherence to treatment, with higher gap days often signaling potential non-adherence.

2.4.1 Modeling on Gap Days

For the modeling of gap days, a clean and focused dataset was created through a series of preprocessing steps to ensure high-quality data. Records were filtered to include only those with drug intake times between the median and upper quartile. Further refinement retained records where the number of people taking the drug also fell within this range. To ensure representation, six second-level ATC codes (A10, J01, L01, N05, N06, S01) were selected, and three drugs were randomly sampled from each. After removing two drugs with significantly higher intake frequencies, a focused dataset of 16 unique drugs was obtained.

Given the substantial proportion of zero values in the gap days, a two-stage hurdle model was used for the analysis. In the first stage, logistic regression was employed to model the probability of having any gap days:

$$\text{logit}(P(\text{gap\_days} > 0)) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_{n-1} X_{n-1},$$

where $X_1, X_2, \ldots, X_{n-1}$ represent one-hot encoded drug types, and one drug serves as the reference category.

In the second stage, conditional on having non-zero gap days, a truncated negative binomial model was fitted to model the count of gap days:

$$\log(E(\text{gap\_days}|\text{gap\_days} > 0)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n-1} X_{n-1} + \log(\text{duration}),$$

where duration (calculated as $|\text{drug\_era\_end\_date} - \text{drug\_era\_start\_date} + 1|$) serves as an offset term. The truncated negative binomial model accommodates overdispersion with variance defined as $\text{Var}(Y) = \mu + \alpha\mu^2$, where μ is the predicted mean, and α was fixed at 1.0. The final expected gap days were calculated as:

$$E(\text{gap\_days}) = P(\text{gap\_days} > 0) \times E(\text{gap\_days}|\text{gap\_days} > 0).$$

2.4.2 Gap Rate

Since gap days are inherently influenced by the total duration of drug use, a normalized measure called the gap rate was defined as:

$$\text{Gap Rate} = \frac{\text{Gap Days}}{\text{Duration}},$$

where duration is calculated as $|\text{drug\_era\_end\_date} - \text{drug\_era\_start\_date} + 1|$. This measure accounts for variations in treatment duration, providing a standardized metric for adherence evaluation.

3 Results

3.1 Exploratory Data Analysis (EDA) on the Dataset

3.1.1 Drug Information

There are 1,620 different drugs in the dataset. The dataset was processed for further analysis, and the results are summarized in the table below. It shows the intersections of drugs across four datasets: Ingredients, ATC, DrugBank, and ChEBI. Each row represents a unique combination of dataset presence, with the corresponding count shown in the rightmost column. The table presents how many drugs are shared between the datasets and which combinations of inclusion and exclusion occur.

Figure 1 provides a detailed visualization of the intersections between the Ingredients, ATC, DrugBank, and ChEBI datasets. It displays the size of each

Ingredients	ATC	DrugBank	ChEBI	Count
O	O	O	X	4
X	O	O	O	7
X	O	O	X	216
X	O	X	X	204
X	X	O	X	74
X	X	X	O	3
X	X	X	X	1,064

Table 1: Intersections of drugs across Ingredients, ATC, DrugBank, and ChEBI datasets. O indicates presence, and X indicates absence.

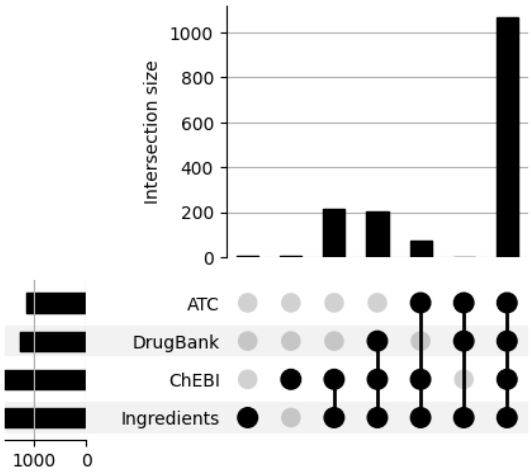


Figure 1: UpSet plot of drugs with Ingredients, ATC Code, DrugBank, and ChEBI Code

intersection, offering a clearer insight into the distribution of shared and unique drugs across the datasets.

From the plots and summarized data, several key observations can be made. ChEBI and Ingredients are almost always present across the datasets. Only four drugs lack ingredient information, and just ten drugs lack a ChEBI code. This is expected, as substances inherently have ingredients, and most drugs tend to have a corresponding ChEBI code due to their chemical nature. Among all missing properties, the absence of an ATC code is the most common. Below is the table showing cases of missing information and representative examples:

Missing ingredient information occurs when the ingredient is implied in the drug name but not explicitly recorded. Substances like Senna Leaves are general in nature, and their active ingredients are often omitted, presenting dataset limitations in handling such substances. The absence of ChEBI codes arises from dataset errors, lack of defined active ingredients (e.g., Senna pod), or incom-

Reason	Examples
Lack of ATC code	Camphor, choline, coal tar, cod liver oil, dextran 70, sorbate, Influenza A virus, glycogen, etc.
Lack of ChEBI code	Lactobacillus paracasei, senna pod, sodium lauryl phosphate, cannabidiol, ursodiol, prucalopride, Bifidobacterium breve, octenidine, hepatitis A virus strain CR 326F antigen, inactivated11, fish oil (containing omega-3 acids), etc.
Lack of DrugBank information	Alpha 1-antitrypsin, clopamide, cyclofenil, dextran 70, chorionic gonadotropin, sorbate, Influenza A virus, glycogen, etc.
Lack of ingredient information	Senna leaves, cobalamins, calcium phosphate dibasic, insulin & pork.

Table 2: Cases of missing information and representative examples.

patibility with the ChEBI classification system (e.g., Lactobacillus paracasei). Missing ATC codes are the most frequent issue. Some substances have ATC codes but were omitted due to dataset inaccuracies (e.g., Camphor). General substances and non-therapeutic ingredients (e.g., PPG-1-PEG-9 Lauryl Glycol Ether) are also excluded. Missing DrugBank information often parallels issues with ATC codes. These gaps result from dataset omissions or misalignment with DrugBank's classification framework, as substances without ATC classifications are less likely to appear in DrugBank. Addressing these gaps can improve dataset accuracy and provide a more comprehensive understanding of drug classifications and properties.

3.1.2 ATC Code Analysis on Drugs

The Anatomical Therapeutic Chemical (ATC) classification system organizes drugs into different levels based on their therapeutic use and chemical properties. One drug can have multiple ATC codes, reflecting its diverse functions or applications. Below is the distribution of the number of ATC codes per drug, indicating that most drugs have a single ATC code, while a smaller fraction has multiple codes (up to 10). This pattern reflects the multifunctionality of certain drugs.

A pie chart illustrating the distribution of first-level ATC codes provides additional insights into how drugs are categorized. The dataset spans all 14 first-level ATC codes, with some categories dominating. The largest categories are 'A: Alimentary Tract and Metabolism' and 'N: Nervous System,' accounting for a significant portion of the dataset, as evident from the bar chart and their respective proportions in the pie chart (13.8% and 13.5%). Other notable categories include 'D: Dermatologicals' and 'C: Cardiovascular System,' each contributing over 9% to the total distribution. Smaller categories, such as 'V: Various,' 'H: Systemic Hormonal Preparations,' and 'P: Antiparasitic Products,' form a minor fraction, indicated by smaller pie slices and lower bar chart counts.

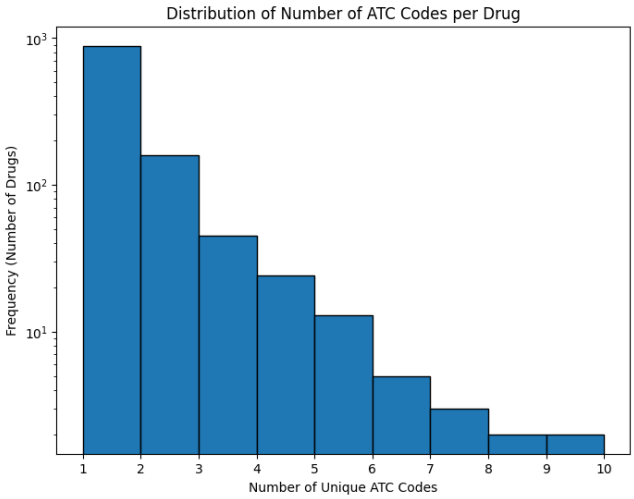


Figure 2: Distribution of number of ATC codes per drug

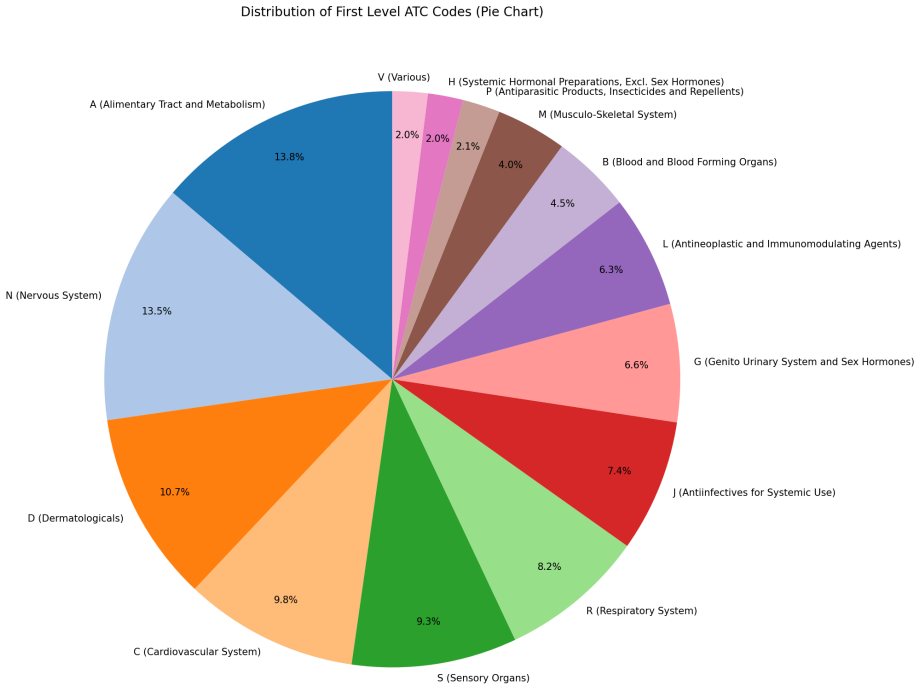


Figure 3: Distribution of 1st level ATC Code

3.1.3 Individual-Level Drug Era

The dataset contains 19.96 million records from 281,690 individuals using 1,620 different drugs. A summary of the data is presented in Table 3, showcasing the total number of drug intake records, unique drugs, drugs mapped to ATC codes, and unique persons.

Name	**Number**
Total number of drug intake records	19,959,413
Number of unique drugs	1,620
Unique drugs mapped to ATC codes	1,141
Number of unique persons	281,690

Table 3: Overview of individual-level drug era data, including the total number of drug intake records, unique drugs, drugs mapped to ATC codes, and unique persons in the dataset.

Table 4 provides descriptive statistics on drug usage. These statistics showcase the variability in drug usage across the dataset, including extreme values and distribution patterns.

Statistic	**Times per Drug**	**People per Drug**	**Times per Person & Drug**	**Drugs per Person**
Count	1,620	1,620	5,952,228	281,690
Mean	12,320.625	3,674.215	3.353	21.130
Std	48,440.126	11,299.933	6.232	19.370
Min	1	1	1	1
25%	13	7	1	5
50%	312.5	121	1	17
75%	4,012.5	1,411.5	3	31
Max	927,199	141,912	155	218

Table 4: Descriptive statistics for drug usage. Columns represent: frequency of times each drug is taken, number of people taking each drug, frequency of drug intake per person and drug, and number of drugs taken by each person. This table presents the variability and distribution of drug usage across the dataset.

The data shows significant variability in drug usage patterns. On average, each drug is taken 12,320.625 times, with some drugs being used as few as once and others nearly a million times (maximum: 927,199). Similarly, the number of people taking a single drug varies widely, with a mean of 3,674.215 individuals per drug, but some drugs are taken by just one person, while others are used by over 140,000 individuals.

For individual drug usage, the number of times each person takes a given drug is often low, with a median of one, reflecting short-term or one-time treatments. However, the maximum count of 155 indicates long-term or recurring drug use for specific drugs. The number of drugs taken per person ranges from

1 to 218, with an average of 21 drugs per person, showing the diversity in individual treatment histories.

This analysis underscores the richness and complexity of the dataset, providing a foundation for further investigations into drug usage trends, adherence patterns, and treatment trajectories.

3.2 Drug Switch

3.2.1 Overview and Exploratory Data Analysis (EDA) of Drug Switches

The dataset provides detailed information about drug switches derived from drug eras. Table 5 summarizes the key statistics related to drug eras and switches, organized by type of switch and ATC levels. The values may underestimate the actual counts, as not all drugs are mapped to ATC codes. Switches within the same ATC levels reflect transitions among related therapeutic agents, while switches across different levels indicate broader treatment changes.

Category	**Count**	**Percentage**
Drug Eras		
Total drug eras (original dataset)	19,959,413	
Closest subsequent drug eras	18,926,988	94.82%
Multiple closest subsequent drug eras	8,694,698	43.56%
Drug Switches		
Total switches	37,254,676	
Switches with the same drug	5,978,429	16.05%
Switches with different drugs	31,276,247	83.95%
Switches by ATC Level		
Within same 3rd level ATC code	7,173,001	19.25%
Between different drugs	1,685,028	4.52%
Within same 2nd level ATC code	8,941,387	24.00%
Between different drugs	3,453,414	9.27%
Within same 1st level ATC code	13,504,325	36.25%
Between different drugs	8,016,352	21.52%

Table 5: Summary of drug switch statistics, including total switches, switches with the same and different drugs, and switches grouped by ATC classification levels. ATC levels represent the hierarchical structure of drug classifications, where a lower level (e.g., 3rd) represents finer granularity.

Switch intervals represent the time between two consecutive drug eras for an individual. Table 6 summarizes the statistical distribution of switch intervals across all recorded drug switches. Additionally, Figure 4 provides visualizations of these intervals at varying ranges. From the data, 7.96% of the switches happen within 7 days, and 27.44% occur within 1 month. Half of the switches occur within 1.5 months, and 75% happen within 2.5 months. The cumulative distribution shows that 80.77% of the switches occur within 3 months and 96.49%

happen within 1 year. The histogram illustrates a clear peak at 30 days, which is attributed to the structure of the drug era data (a single gap of over 30 days separates drug eras).

Statistic	**Value**
Count	37,254,676
Mean	85.46 days
Standard Deviation (std)	194.23 days
Minimum	1 day
25%	27 days
50% (Median)	43 days
75%	73 days
Maximum	9,660 days

Table 6: Summary statistics for switch intervals across all drug switches in the dataset. The table describes the count, mean, standard deviation, and percentile values of switch intervals.

Figure 4 displays four histograms that illustrate the distribution of switch intervals for different ranges. The first plot covers the entire range of switch intervals, showing a heavy-tailed distribution with a significant concentration at shorter durations. The second plot focuses on intervals within a year (0–365 days), which account for most of the dataset, with a large proportion of intervals below 100 days. The third plot provides finer granularity for intervals up to 3 months (0–90 days), revealing peaks at regular intervals likely related to treatment schedules. Finally, the fourth plot shows the shortest durations (0–30 days), with a notable spike at 30 days due to the structural constraint of the dataset (gap ¿ 30 days would separate the drug era). The visualizations and summary statistics together provide a comprehensive understanding of switch intervals, revealing both common patterns and structural properties of the data.

The dataset contains 442,781 unique drug switch patterns. Table 7 summarizes the descriptive statistics for these switch patterns in terms of their frequency by times and by the number of people involved. Most switch patterns are not common, but a few switches are highly frequent.

Table 8 presents the top 5 most frequent switch patterns within the same N06 drugs (psychoanaleptics) and between different 66 drugs. This comparison showcases patterns of continued treatment and therapeutic transitions.

3.2.2 PMI Score on Each Drug Switch Pattern

After filtering out uncommon switch patterns (less than 30 individuals per switch), I identified 76,152 distinct drug switch patterns. Among these, two histograms were constructed: one includes all switch patterns, and the other excludes self-switch patterns (switches involving the same drug). The histograms reveal that self-switch patterns have the highest PMI scores, which aligns with the intuitive observation that patients tend to continue using the same drug.

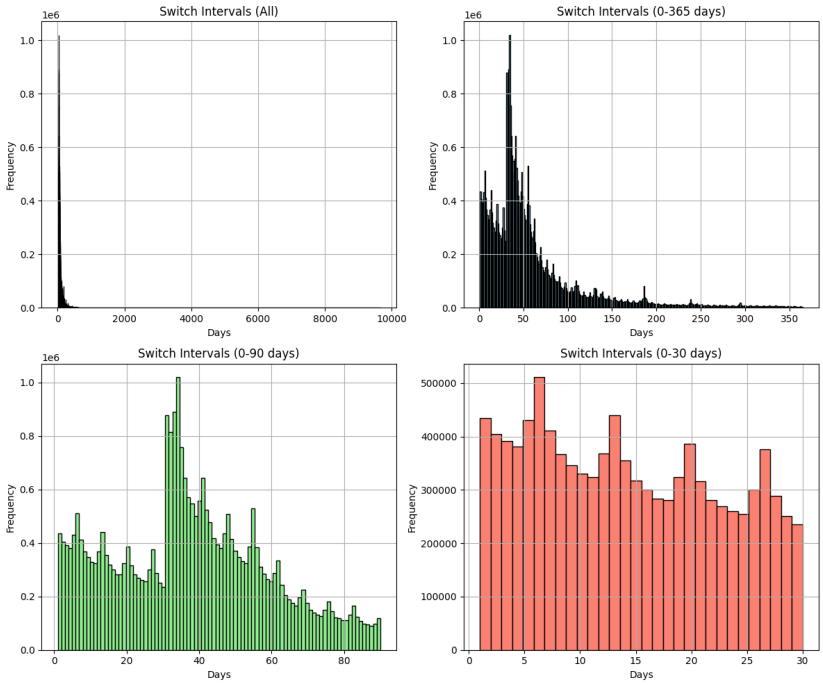


Figure 4: Histograms of switch intervals. The top-left plot shows all intervals, while the others focus on specific ranges (0–365 days, 0–90 days, and 0–30 days). The histograms reveal the patterns of drug switches and the structural features of the dataset.

Statistic	**By Times**	**By People**
Count	442,781	442,781
Mean	84.1	48.1
Std	1,354.9	324.3
Min	1	1
25%	1	1
50%	5	4
75%	21	15
Max	541,893	38,923

Table 7: Summary statistics for drug switch frequencies, showing the times each switch occurred and the number of people involved.

From the statistics in Table 9, approximately half of the switch patterns (25%-75%) exhibit a PMI score near 0, indicating that these transitions are largely random and occur by chance. This observation corresponds to the peak at 0 in the histogram. However, from the histogram we can also see a considerable amount of switch patterns with PMI scores greater than 5, suggesting a

Source Drug	Destination Drug	People Count	Avg. Interval (Days)
Same N06 Drugs			
Amitriptyline	Amitriptyline	15,261	54.3
Citalopram	Citalopram	8,507	59.7
Fluoxetine	Fluoxetine	6,797	72.9
Dothiepin	Dothiepin	3,780	57.9
Paroxetine	Paroxetine	2,829	67.1
Different N06 Drugs			
Amitriptyline	Citalopram	1,119	67.4
Citalopram	Amitriptyline	976	60.0
Amitriptyline	Fluoxetine	964	67.7
Fluoxetine	Amitriptyline	881	65.3
Fluoxetine	Citalopram	786	129.5

Table 8: Top 5 frequent switch patterns within the same N06 drugs (Psychoanaleptics) and between different N06 drugs. The table includes people count and average switch interval in days.

strong association between the source and destination drugs. These high-PMI drug switches warrant further analysis.

Statistic	Value
Count	76,152
Mean	0.225454
Std	1.690515
Min	-6.514141
25%	-0.594804
50%	-0.029662
75%	0.619119
Max	19.286070

Table 9: Summary statistics for PMI scores of drug switch patterns.

The two histograms shown in Figure 5 compare different sets of drug switch patterns. The left histogram represents all 76,152 drug switch patterns, while the right histogram excludes self-switch patterns. Both histograms show a concentration of PMI scores near 0, with a tail extending to higher PMI values. The key difference lies in the rightmost part of the distributions, where the left histogram includes higher PMI scores. This difference arises because self-switch patterns, where individuals continue using the same drug, are more likely to have exceptionally high PMI values.

Example 1: Lorazepam (`N05BA06`) to Quetiapine (`N05AH04`) The switch from Lorazepam, a benzodiazepine for anxiety and insomnia, to Quetiapine, an atypical antipsychotic, has a PMI score of 5.717. This switch occurred 44

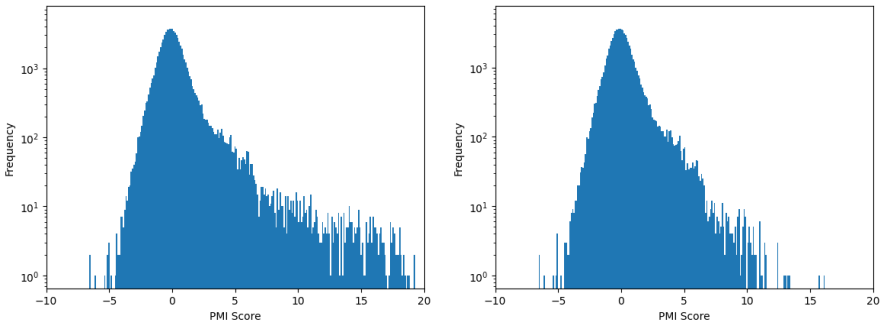


Figure 5: PMI score histograms for drug switch patterns. **Left:** All switch patterns. **Right:** Excluding self-switch patterns.

times, while Lorazepam appeared as a source drug 5,445 times and Quetiapine as a destination drug 5,305 times. This pattern reflects a clinically significant transition, likely for managing comorbid psychiatric conditions or minimizing dependence on benzodiazepines.

Example 2: Oxycodone (`N02AA05`) to Naloxone (`A06AH04`, `V03AB15`) The switch from Oxycodone, an opioid for pain relief, to Naloxone, used to counter opioid overdoses, has a remarkably high PMI score of 11.522. This switch occurred 105 times, with Oxycodone as the source drug 4,910 times and Naloxone as the destination drug 251 times. This strong association likely reflects interventions for opioid addiction or overdose management. Interestingly, the reverse switch, Naloxone to Oxycodone, also shows a high PMI score of 11.255, further implying the complexity of opioid-related treatments.

These drug switches, characterized by high PMI scores, offer valuable insights into treatment pathways and can be prioritized for deeper exploration to understand their clinical and pharmacological significance.

3.2.3 PMI Score Between Drug Switch Patterns

After filtering out uncommon switches (occurring in fewer than 30 individuals) and keeping only N-drugs, I identified 1,875 unique drug switch patterns. Using these patterns, I generated combinations of drug switch pairs, resulting in a total of 1,756,875 pairs. For each pair of drug switches, the PMI score was calculated.

The results presented here include only the 635,087 drug switch pairs with non-zero co-occurrence. The summary statistics of the PMI scores for these pairs are shown in Table 10. The mean and median PMI scores are both greater than 2, with the histogram peaking at approximately 2. This indicates that drug switch patterns, in general, exhibit positive correlations.

Figure 6 shows the histogram of PMI scores for these drug switch pairs. The distribution demonstrates a peak around 2, suggesting that drug switch pairs

Statistic	Value
Count	635,087
Mean	2.548083
Std	1.807296
Min	-4.132932
25%	1.239288
50%	2.367151
75%	3.717124
Max	11.878805

Table 10: Summary statistics for PMI scores between drug switch pairs.

tend to co-occur more frequently than would be expected by chance.

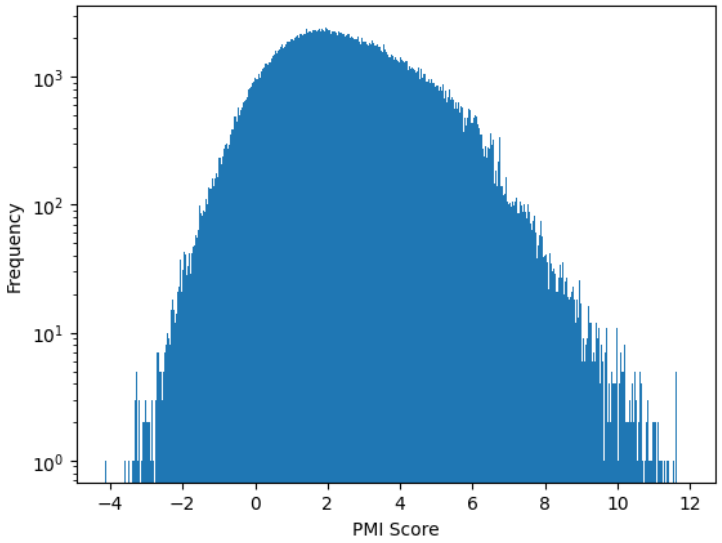


Figure 6: Histogram of PMI scores for drug switch pairs. The peak around 2 indicates positive correlations between most drug switch pairs.

Example 1: Lorazepam (`N05BA06`) → Quetiapine (`N05AH04`) and Quetiapine (`N05AH04`) → Lamotrigine (`N03AX09`) This pair of drug switches has a strong PMI score of 8.7857, indicating that these transitions co-occur far more frequently than expected by chance. Lorazepam, commonly used for anxiety or agitation, transitions to Quetiapine, a mood stabilizer, which is then followed by Lamotrigine to further stabilize mood or address bipolar disorder. This sequence reflects common psychiatric treatment strategies for managing comorbidities and minimizing side effects.

Example 2: Caffeine (N06BC) → Pizotiline (N02CA03) and Amitriptyline (N06AA09) → Ergotamine (N02CA02) This pair of switches exhibits a high PMI score of 9.831, reflecting a strong clinical focus on migraine or headache management. Caffeine, used for general pain relief, transitions to Pizotiline for migraine prevention, while Amitriptyline, commonly prescribed for chronic migraines, transitions to Ergotamine for acute migraine treatment. These co-occurrences indicates typical treatment pathways for managing both acute and long-term migraine conditions.

These results underscore the importance of identifying and analyzing high-PMI drug switch pairs, as they provide valuable insights into treatment trajectories and clinical practices. These patterns can guide personalized medicine by helping clinicians understand the connections between different therapies and patient outcomes.

3.2.4 Clustering on Drug Switch Patterns with N-drugs

The network visualization in Figure 7 represents drug switch patterns in our dataset. In this network, nodes correspond to individual drug switches, and edges connect switches that frequently co-occur, as measured by PMI scores above 4.0. Each node is assigned a color based on its community membership, identified using the Louvain community detection algorithm.

The visualization reveals distinct clusters of related drug switches, suggesting common patterns in medication changes. The network consists of 1860 nodes and 133,506 edges, which are organized into 9 distinct communities. These clusters indicate that certain drug switches frequently occur together, far more often than would be expected by chance. This behavior may reflect shared therapeutic strategies or related treatment pathways. For example, switches within the same community may involve drugs used to treat the same condition or drugs that are commonly used in sequential therapies.

The varying sizes and densities of the communities reflect differences in the scale and nature of the switching patterns. Some communities are tightly connected, suggesting highly specific patterns of co-occurrence, while others are larger and more diffuse, indicating broader associations. The network structure uncovers a hierarchical organization of drug switches, with some clusters representing specialized treatment regimens and others capturing general trends across broader drug categories.

3.3 Gap Days

3.3.1 Focused Dataset

The focused dataset consists of drug intake records filtered to include specific drugs and their usage information. An overview of the dataset is provided in Table 11, including the total number of drug intake records, unique drugs, and unique patients.

Table 12 provides detailed information about the drugs included in the focused dataset. For each drug, the table lists the drug's name, its corresponding

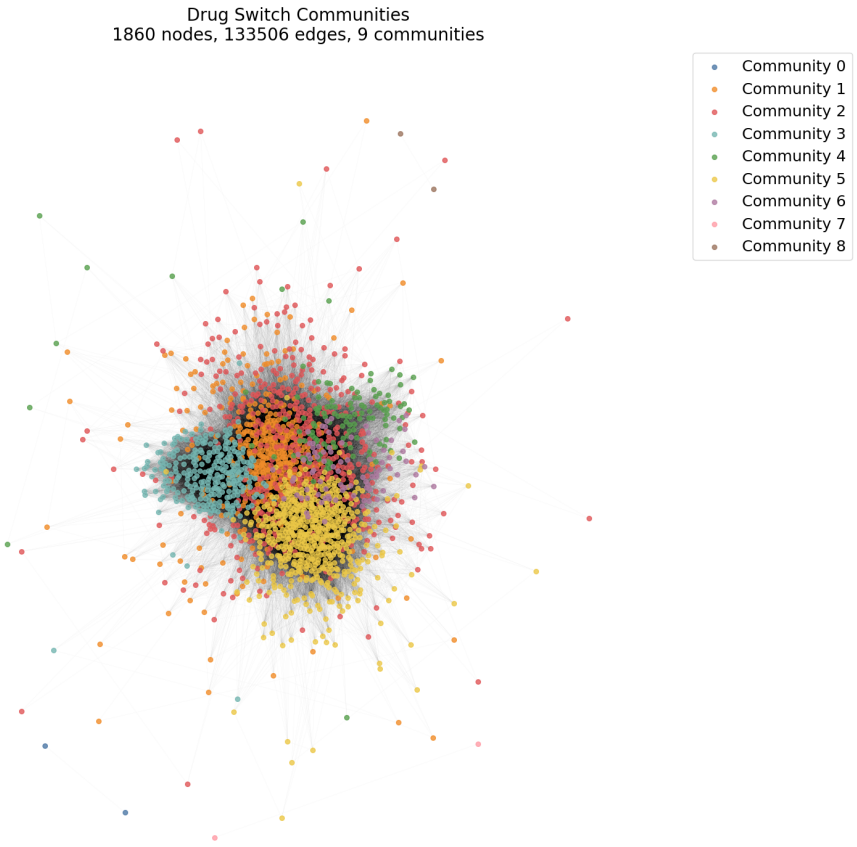


Figure 7: Network visualization of drug switch patterns. Nodes represent individual drug switches, and edges connect frequently co-occurring switches. Each color denotes a community detected by the Louvain algorithm. The network contains 1860 nodes, 133,506 edges, and 9 distinct communities.

Name	**Number**
Total number of drug intake records	17,668
Number of unique drugs	16
Number of unique persons	8,117

Table 11: Overview of the focused dataset.

ATC code(s), the total number of times the drug was taken, and the number of patients who used the drug.

The dataset demonstrates that the frequency of drug intake and the number of patients for each drug are relatively consistent, with no extreme outliers observed. While there are slight variations, the data is uniformly distributed across the included drugs, ensuring that no single drug disproportionately dominates the dataset. This uniformity suggests that the selection process was effective in maintaining balance, making the dataset suitable for further analysis without bias toward any specific drug.

Drug Concept Name	**ATC Code(s)**	**Times Taken**	**Patients**
Exenatide	A10BJ01	1,821	1,219
Linagliptin	A10BH05	1,751	1,157
Vildagliptin	A10BH02	1,648	1,039
Cefuroxime	J01DC02, S01AA27	1,496	908
Levofloxacin	J01MA12, S01AE05	1,337	835
Chlortetracycline	J01AA03, S01AA02	1,322	757
Tobramycin	J01GB01, S01AA12	1,294	629
Eflornithine	L01XX79	1,124	497
Hydroxyurea	L01XX05	1,102	430
Mercaptopurine	L01BB02	1,030	388
Midazolam	N05CD08	911	368
Methotrimeprazine	N05AA02	626	213
Modafinil	N06BA07	625	177
Rivastigmine	N06DA03	592	159
Acetazolamide	S01EC01	548	144
Lodoxamide	S01GX05	441	125

Table 12: All drugs in the focused dataset, including their ATC codes, number of times taken, and number of patients.

3.3.2 Modeling on Gap Days

A negative binomial model was employed to analyze non-zero gap days, using acetazolamide as the reference drug. The model demonstrated a reasonable fit with an Akaike Information Criterion (AIC) of 40,442 and a pseudo R-squared value of 0.1109, explaining approximately 11.1% of the variance in non-zero gap days. The intercept ($\beta = -1.213$) represents the expected log gap days for acetazolamide, accounting for the duration offset.

Table 13 provides the estimated coefficients for each drug. Positive coefficients indicate an increase in expected gap days, while negative coefficients represent a reduction relative to acetazolamide.

Methotrimeprazine exhibited the strongest positive effect, increasing expected gap days by 61.9% ($\beta = 0.482$, $p < 0.001$), while Linagliptin showed the strongest negative effect, reducing gap days by 62.9% ($\beta = -0.991$, $p < 0.001$).

Drug Name	Coefficient (β)	Std. Error	Effect on Gap Days
Intercept	-1.213	0.065	Reference
Cefuroxime	0.241	0.125	+27.2%
Chlortetracycline	-0.444	0.259	-35.9%
Eflornithine	-0.252	0.089	-22.3%
Exenatide	-0.470	0.075	-37.5%
Hydroxyurea	-0.236	0.085	-21.0%
Levofloxacin	0.181	0.122	+19.8%
Linagliptin	-0.991	0.076	-62.9%
Lodoxamide	-0.399	0.085	-32.8%
Mercaptopurine	-0.308	0.089	-26.5%
Methotrimeprazine	0.482	0.104	+61.9%
Midazolam	-0.620	0.141	-46.2%
Modafinil	0.015	0.081	+1.5%
Rivastigmine	-0.168	0.087	-15.4%
Tobramycin	-0.528	0.197	-41.0%
Vildagliptin	-0.287	0.084	-24.9%

Table 13: Coefficients from the negative binomial model for gap days. Effects are calculated relative to acetazolamide.

Significant reductions in gap days were also observed for Exenatide ($\beta = -0.470$, $p < 0.001$), Midazolam ($\beta = -0.620$, $p < 0.001$), and Tobramycin ($\beta = -0.528$, $p = 0.007$). These results suggest substantial variation in medication adherence patterns across different drugs.

Figure 8 illustrates key model diagnostics. Panel (a) in Figure 8 displays the probabilities of having zero gap days for different drugs. The variation observed in these probabilities reflects differences in adherence levels or prescribing practices among the medications. These variations suggest that certain drugs are more likely to be associated with continuous usage without gaps, while others might be prone to interruptions in treatment. Panel (b) illustrates the relationship between predicted and observed gap days. While precise predictions remain challenging, the plot reveals that the model performs in an unbiased manner. The points are symmetrically distributed around the $y = x$ line, indicating no systematic tendency toward over- or under-prediction. This symmetry demonstrates the reliability of the model in capturing the overall patterns in the data without introducing significant bias.

The negative binomial model reveals significant variations in gap day patterns across drugs, with some medications strongly reducing gap days while others increase them. The model's diagnostics confirm a good fit, providing valuable insights into adherence patterns and informing clinical strategies to optimize medication use.

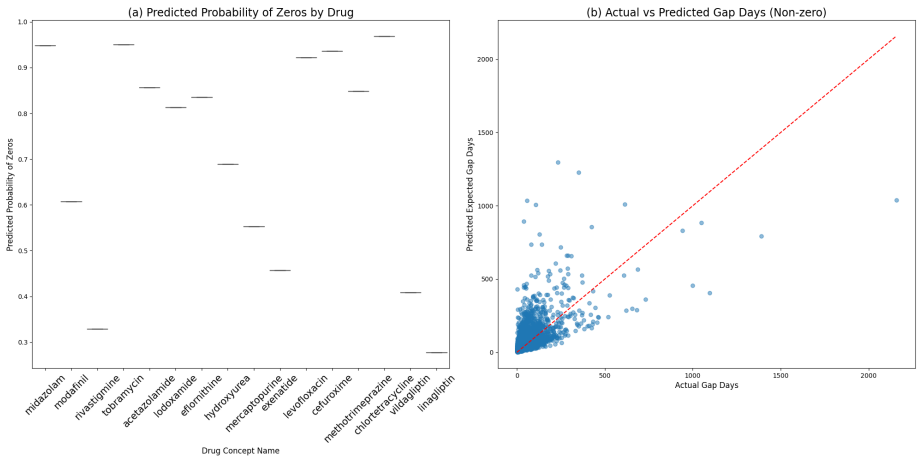


Figure 8: Model diagnostics for gap days. (a) Probabilities of zero gap days across drugs. (b) Predicted vs. observed gap days, showing unbiased predictions around the $y = x$ line.

3.3.3 Gap Rate

The gap rate, defined as the proportion of gap days within a drug era, was analyzed across all drug eras and separately for drug eras involving N06 drugs (psychoanaleptics). Table 14 summarizes the descriptive statistics for gap rates in both groups.

Group	Mean	Std	25%	50%	75%	Min	Max
All Drug Eras	0.064	0.161	0.000	0.000	0.000	0.000	0.999
N06 Drug Eras	0.302	0.242	0.098	0.242	0.459	0.0002	0.999

Table 14: Summary statistics for gap rates in all drug eras and N06 drug eras.

The gap rate for all drug eras exhibits a mean of 0.064, with a substantial proportion of drug eras having a gap rate of 0, as shown by the median and 25th percentile values. This indicates that most drug eras have no interruptions in usage. In contrast, the gap rate for N06 drug eras shows no cases with a 0 gap rate, with a mean of 0.302 and higher variability (standard deviation = 0.242). The minimum gap rate for N06 drug eras is 0.000187, emphasizing the consistent presence of gap days in this category. The observed difference between these two groups may reflect the therapeutic nature of N06 drugs, which often require continuous use for long-term treatment.

Figure 9 provides a side-by-side comparison of histograms for gap rates across all drug eras and N06 drug eras. The left panel shows the distribution of gap rates for all drug eras, with a pronounced peak at 0, representing drug eras with no interruptions. This distribution suggests that uninterrupted drug usage is common for most drugs. The right panel shows the gap rate distribution

for N06 drug eras, where no peak at 0 is observed. Instead, the gap rates are more evenly distributed, with a concentration around a mean value of 0.302. This difference indicates the distinct usage patterns associated with N06 drugs, which require continuous adherence and tend to have fewer interruptions.

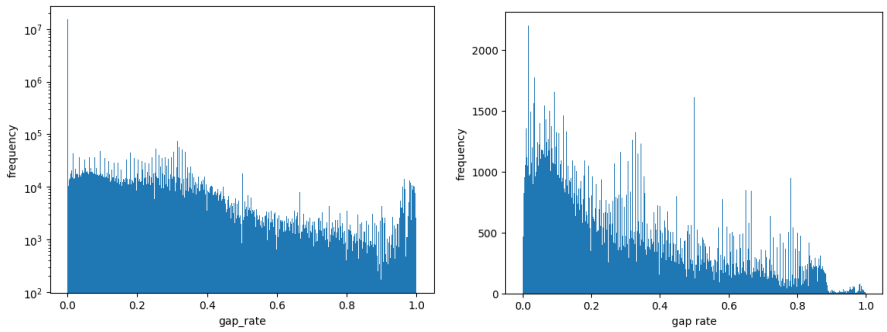


Figure 9: Histograms of gap rates for all drug eras (left) and N06 drug eras (right). The distribution for all drug eras shows a prominent peak at 0, while N06 drug eras exhibit consistently non-zero gap rates.

In addition to the general distributions, Figure 26 presents histograms of the gap rates for four types of antidepressants of N06 drugs. These histograms illustrate the gap rate patterns for four specific N06 drugs, providing additional granularity. The first histogram corresponds to trazodone, showing a concentration of gap rates around 0.25. The second histogram represents amitriptyline, which demonstrates a wider spread of gap rates, peaking at approximately 0.4. The third and fourth histograms correspond to the drugs fluoxetine and sertraline, respectively, each showing distinct patterns that reflect unique adherence trends and prescribing practices for these medications. The variations across the four histograms reinforce the idea that even within drugs of the same therapeutic area, there are distinct patterns in gap rates based on the specific drug.

Overall, the observations demonstrate the importance of drug-specific characteristics in influencing gap rates and underscores the utility of further exploring these metrics to inform strategies for improving adherence and treatment outcomes.

4 Discussion

This study provides a comprehensive analysis of drug utilization patterns and their implications for understanding population-level adherence behaviors, treatment switches, and potential genetic underpinnings. The findings underscore the complexity and variability of drug-related data, necessitating careful curation and contextual interpretation for effective modeling.

The dataset exhibits substantial variability across patient demographics, treatment durations, and drug-switching behaviors. This variability empha-

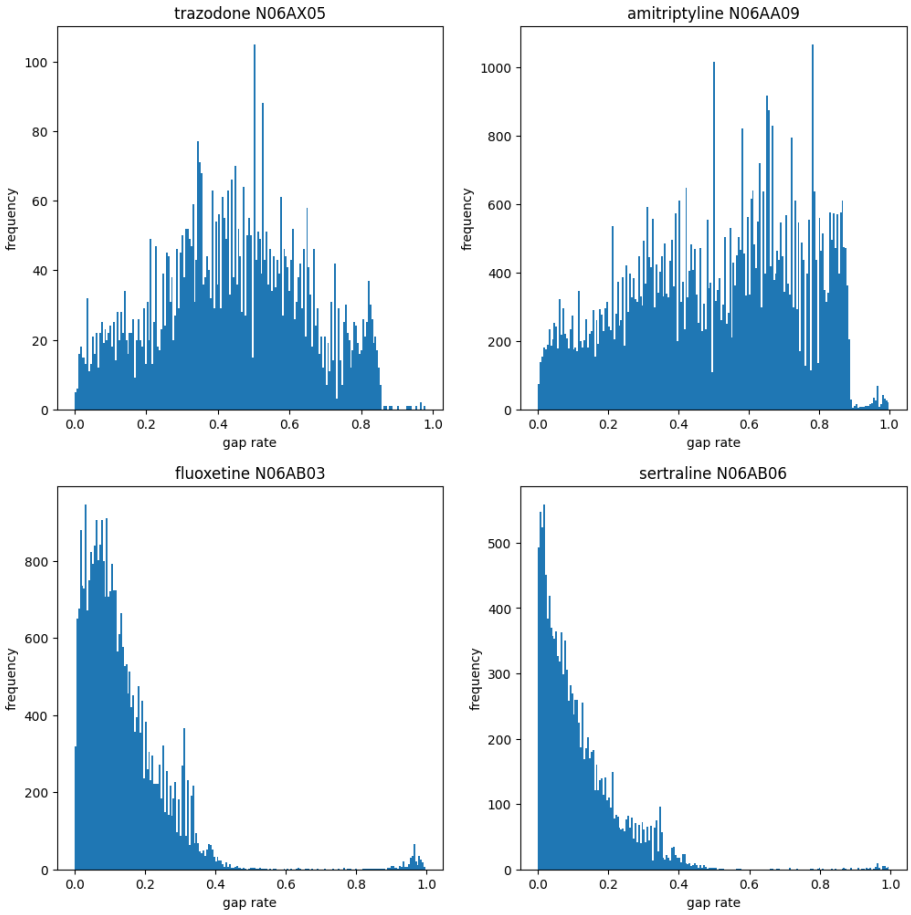


Figure 10: Histograms of gap rates for four N06 drugs. Each drug exhibits distinct gap rate patterns, reflecting differences in adherence and prescribing practices.

sizes the importance of rigorous data filtering to ensure meaningful insights can be extracted for modeling. For instance, the frequent absence of ATC codes and the variable quality of drug exposure records pose challenges to creating comprehensive and interpretable relationships between drug utilization patterns and genetic markers. Addressing these gaps by incorporating additional phenotypic and genotypic data will enhance the utility and accuracy of the analysis.

The analysis of drug-switching patterns presents an opportunity to infer treatment efficacy and side effects indirectly. However, while the computed PMI scores reveal statistically significant associations between certain drug switches, interpreting these transitions is not straightforward. Many switches may result from clinical decisions based on side effects or disease progression, but the lack of explicit information in the dataset limits the depth of analysis. Further integration of clinical outcome data and patient-reported experiences is necessary to better understand the drivers of switching behaviors.

The modeling of gap days as a measure of adherence demonstrated its ability to capture drug-specific patterns, with the logistic regression model showing strong statistical significance. However, the moderate pseudo R-squared value (11.09%) suggests that drug type alone cannot fully explain the observed variance. Other unmodeled factors, such as socioeconomic status, patient comorbidities, or pharmacogenomic interactions, likely play a role. As such, the current model is more suitable for population-level analyses rather than precise individual predictions, emphasizing the need for richer datasets to enable more granular insights.

Despite the progress achieved, several limitations of the dataset constrain the study's conclusions. The absence of detailed information on patients' genetic profiles, lifestyle factors, and clinical outcomes restricts deeper exploration of the observed drug utilization patterns. Additionally, while PMI-based clustering of drug switches identifies potentially meaningful groupings, their biological and clinical relevance remains unclear without corroborating evidence from other domains, such as pharmacogenomics or clinical trials.

Future research should aim to integrate this dataset with whole-genome sequencing data, such as the UK Biobank WGS, to investigate genotype-phenotype associations in drug adherence and switching. By linking drug utilization patterns with genetic markers and phenotypic traits, we can deepen our understanding of the interplay between genetic predispositions and treatment responses. Additionally, leveraging disease trajectory methodologies, as demonstrated in studies like the Danish Disease Trajectory Browser, could provide further insights into temporal patterns of drug switching and adherence [2, 5].

In conclusion, this study illustrates both the potential and the challenges of using large-scale drug utilization data to explore treatment adherence and switching behaviors. While the current analyses establish a strong foundation for population-level insights, future integration with genetic and clinical datasets is essential to fully realize the promise of personalized medicine.

References

[1] Bjarni V. Halldorsson, Hannes P. Eggertsson, Kristjan H. S. Moore, et al. "The sequences of 150,119 genomes in the UK Biobank". In: *Nature* 607.732 (2022), pp. 732–740. DOI: 10.1038/s41586-022-04965-x.

[2] Amalie D. Haue, Jose J. Almagro Armenteros, Peter C. Holm, et al. "Temporal patterns of multi-morbidity in 570,157 ischemic heart disease patients: A nationwide cohort study". In: *Cardiovascular Diabetology* 21.87 (2022). DOI: 10.1186/s12933-022-01527-3.

[3] Tuomo Kiiskinen, Pyry Helkkula, Kristi Krebs, et al. "Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases". In: *Nature Medicine* 29.1 (2023), pp. 209–218. DOI: 10.1038/s41591-022-02122-5.

[4] OHDSI. *Drug Era Documentation*. Accessed on January 27, 2025. n.d. URL: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:drug_era (visited on 01/27/2025).

[5] Troels Siggaard et al. "Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients". In: *Nature Communications* 11.1 (2020), p. 4952. DOI: 10.1038/s41467-020-18682-4.

Appendices

2nd-level ATC Codes

ATC Code	Category	Number of Second-Level Codes	Presence	Absence	Presence Percentage
A	Alimentary Tract and Metabolism	16	13	3	81.25%
B	Blood and Blood Forming Organs	5	4	1	80.00%
C	Cardiovascular System	9	9	0	100.00%
D	Dermatologicals	11	11	0	100.00%
G	Genito Urinary System and Sex Hormones	4	4	0	100.00%
H	Systemic Hormonal Preparations, Excl. Sex Hormones	5	5	0	100.00%
J	Antiinfectives for Systemic Use	6	6	0	100.00%
L	Antineoplastic and Immunomodulating Agents	4	4	0	100.00%
M	Musculo-Skeletal System	6	5	1	83.33%
N	Nervous System	7	7	0	100.00%
P	Antiparasitic Products, Insecticides and Repellents	3	3	0	100.00%
R	Respiratory System	6	5	1	83.33%
S	Sensory Organs	3	3	0	100.00%
V	Various	9	5	4	55.56%
SUM		94	84	10	89.36%

Figure 11: Statistics on the presence of drugs categorized by second-level ATC codes. The table shows that most ATC categories have complete or near-complete drug representation in the dataset, with many categories (e.g., Cardiovascular System, Dermatologicals) showing 100% presence. However, certain categories like "Alimentary Tract and Metabolism" and "Various" have lower representation, with the presence percentages dropping to 81.25% and 55.56%, respectively.

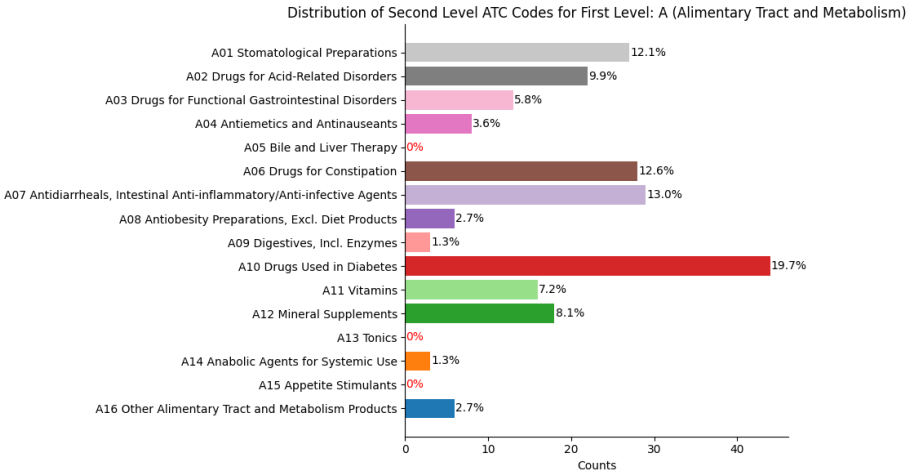


Figure 12: Distribution of second level ATC Codes for first level A: Alimentary Tract and Metabolism

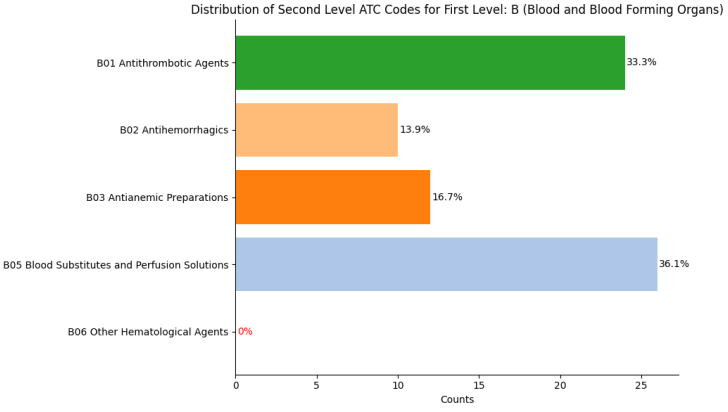


Figure 13: Distribution of second level ATC Codes for first level B: Blood and Blood Forming Organs

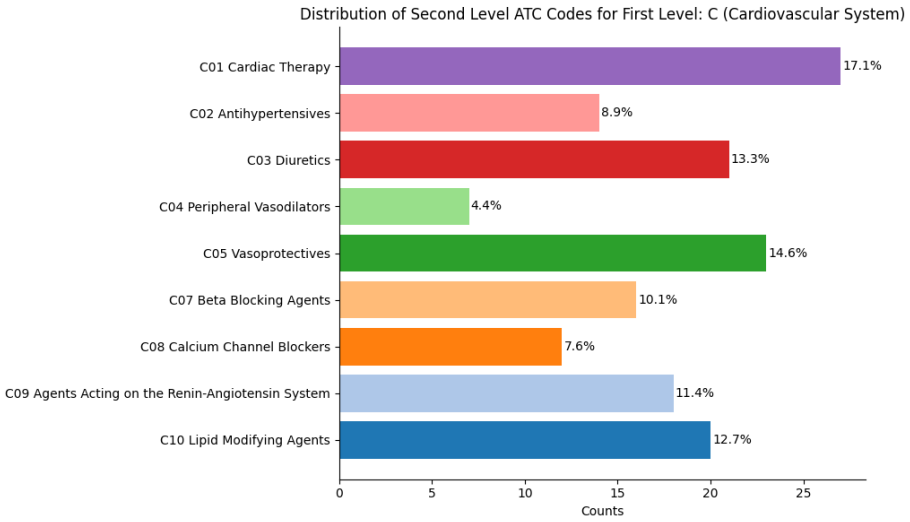


Figure 14: Distribution of second level ATC Codes for first level C: Cardiovascular System

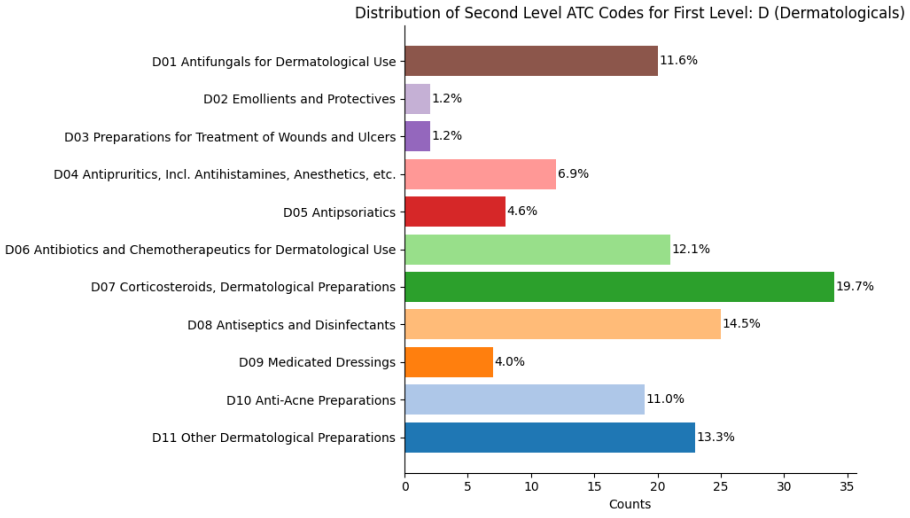


Figure 15: Distribution of second level ATC Codes for first level D: Dermatologicals

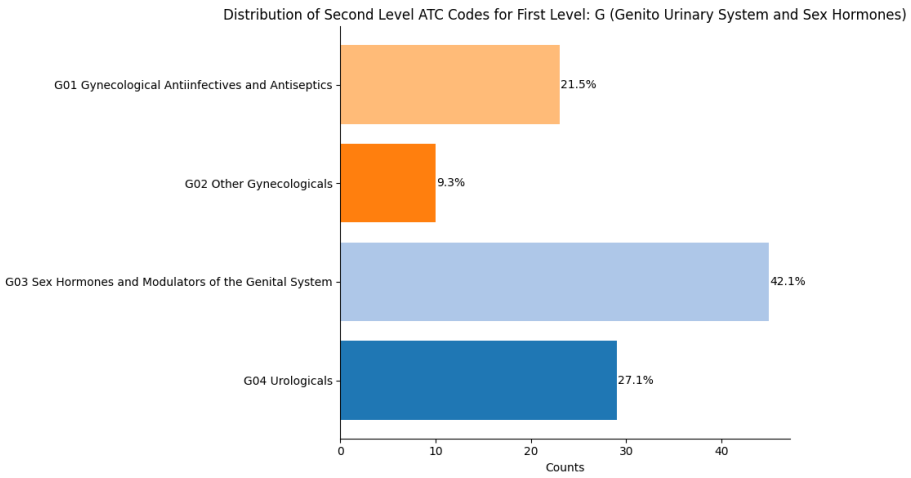


Figure 16: Distribution of second level ATC Codes for first level G: Genito Urinary System and Sex Hormones

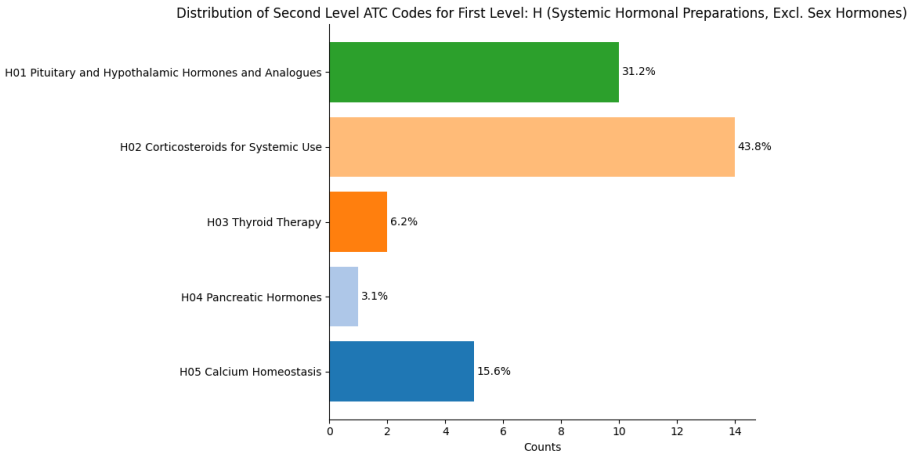


Figure 17: Distribution of second level ATC Codes for first level H: Systemic Hormonal Preparations, Excl. Sex Hormones

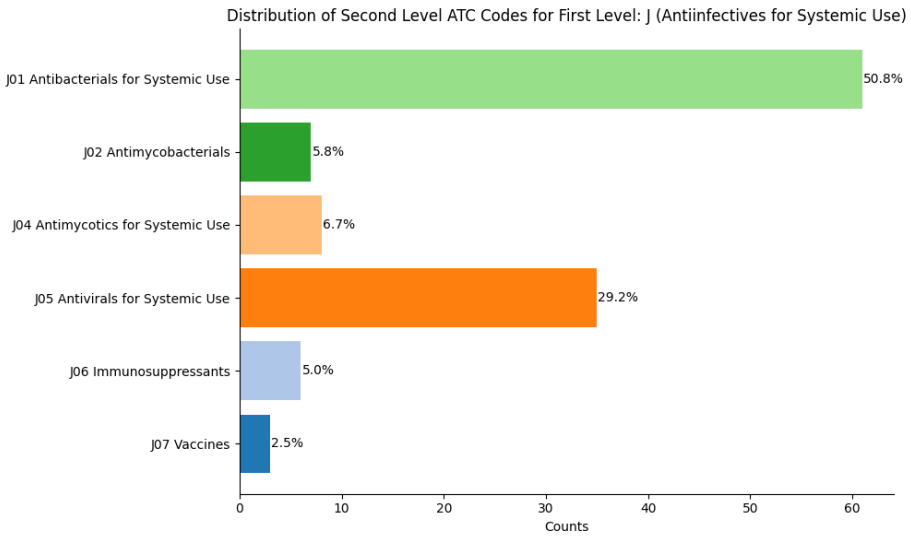


Figure 18: Distribution of second level ATC Codes for first level J: Antiinfectives for Systemic Use.

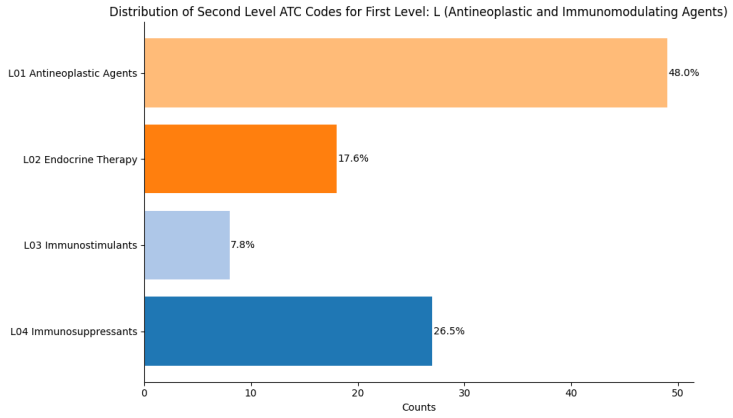


Figure 19: Distribution of second level ATC Codes for first level L: Antineoplastic and Immunomodulating Agents

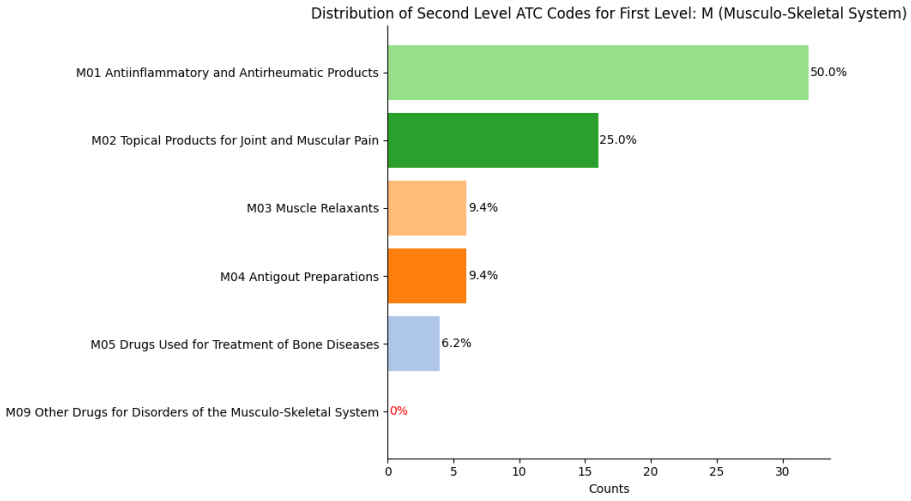


Figure 20: Distribution of second level ATC Codes for first level M: Musculo-Skeletal System

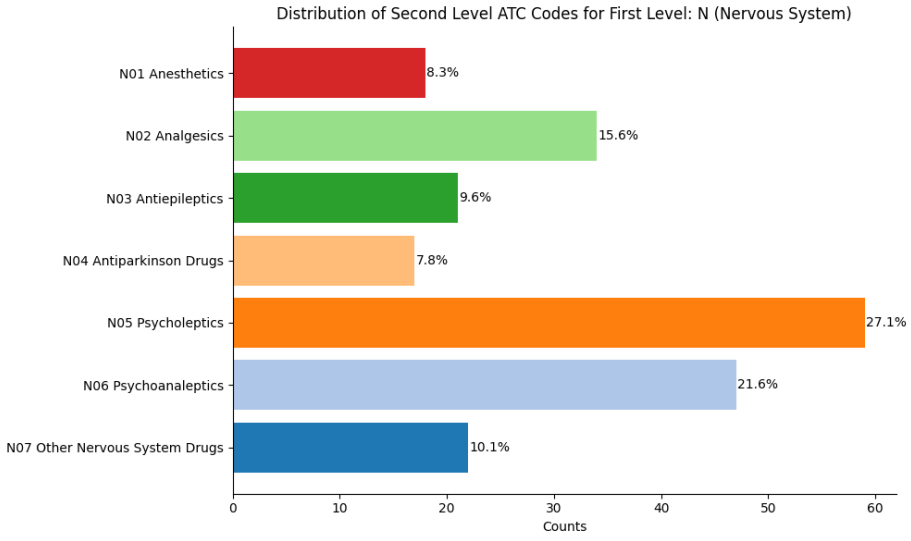


Figure 21: Distribution of second level ATC Codes for first level N: Nervous System

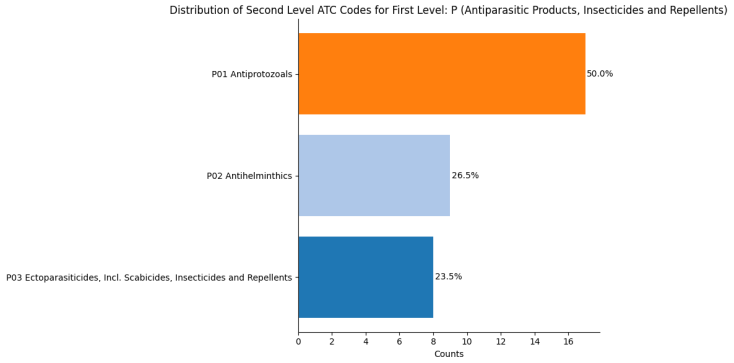


Figure 22: Distribution of second level ATC Codes for first level P: Antiparasitic Products, Insecticides and Repellents

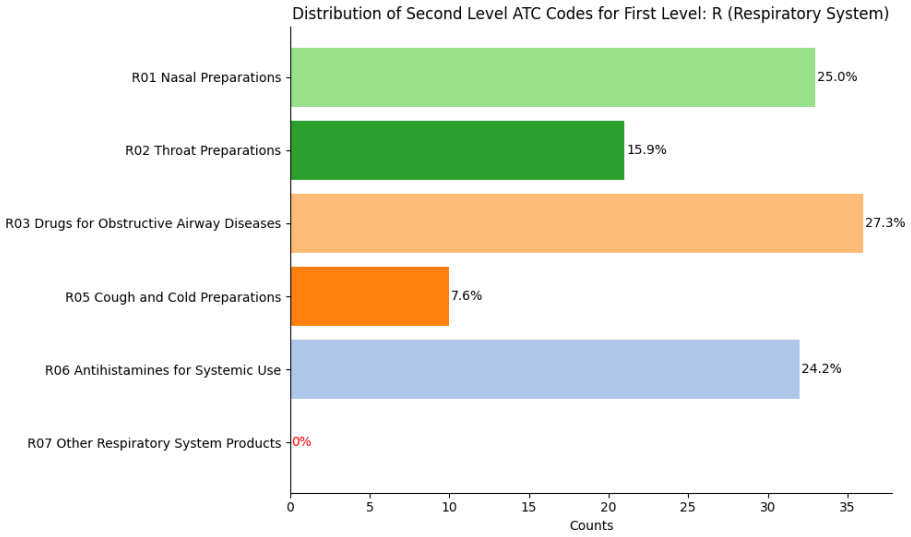


Figure 23: Distribution of second level ATC Codes for first level R: Respiratory System

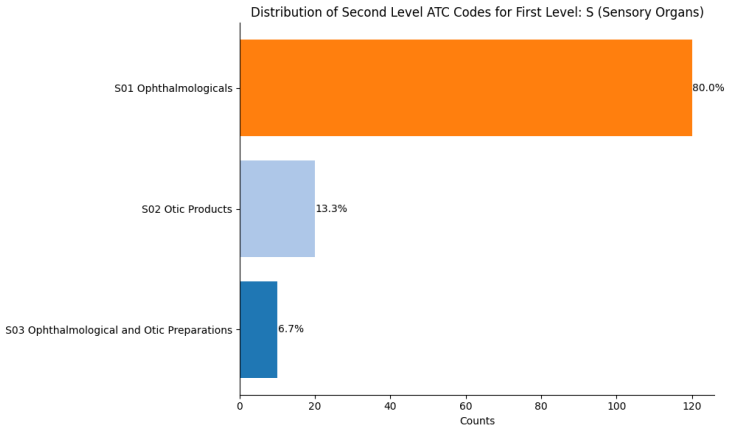


Figure 24: Distribution of second level ATC Codes for first level S: Sensory Organs

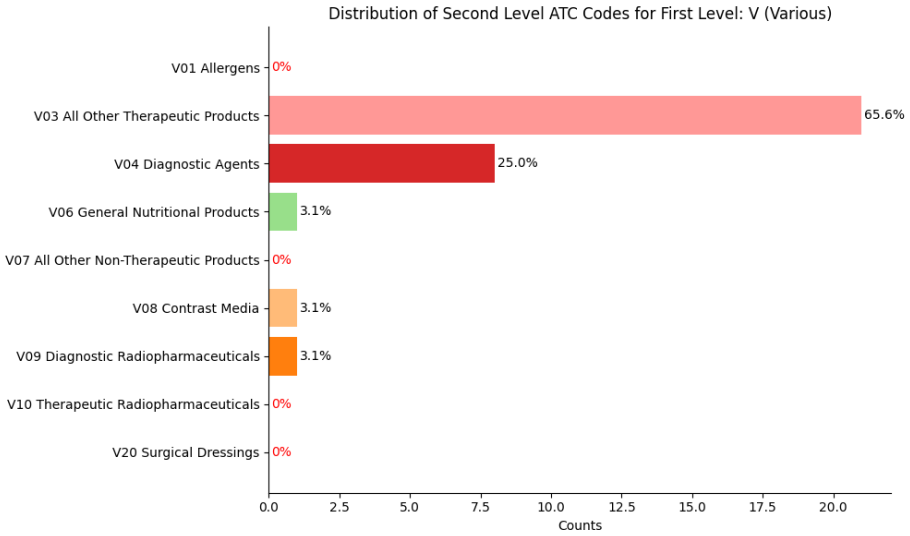


Figure 25: Distribution of second level ATC Codes for first level V: Various

Histograms for individual-level drug era EDA

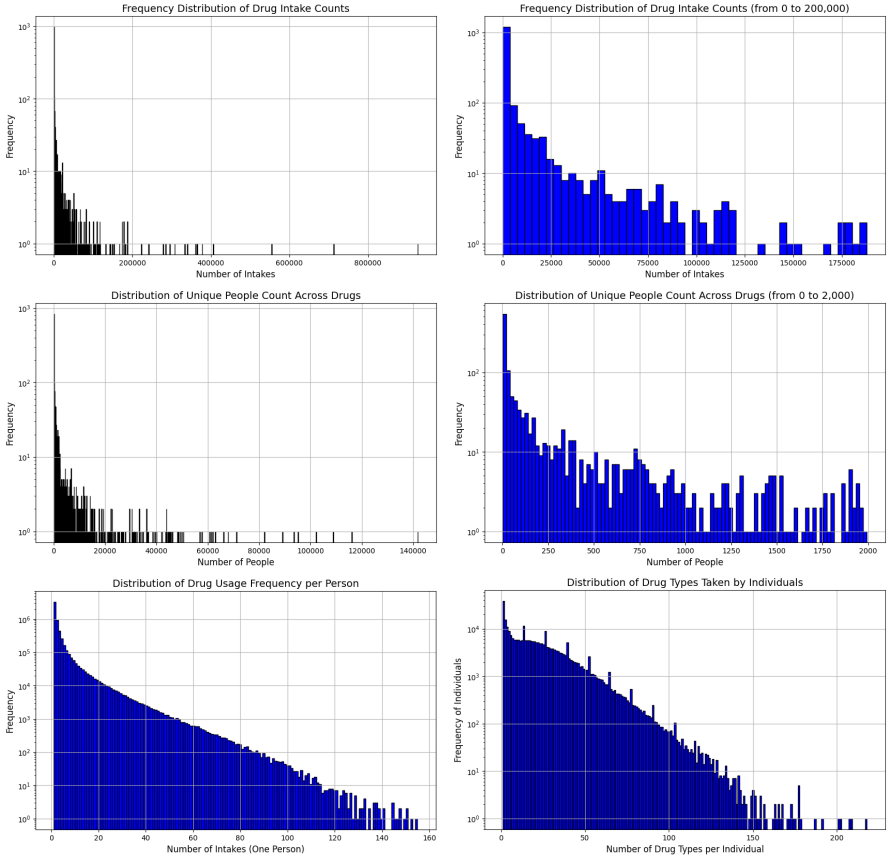


Figure 26: Series of histograms illustrating various aspects of drug intake patterns in the dataset. The top-left panel shows the frequency distribution of drug intake counts across all drugs, while the top-right panel focuses on drug intake counts within a restricted range (0 to 200,000). The middle-left panel displays the distribution of unique people counts for each drug, and the middle-right panel presents the same distribution but limited to a range (0 to 2,000). The bottom-left panel depicts the distribution of drug usage frequency per individual, and the bottom-right panel represents the distribution of the number of drug types taken by individuals. Each histogram provides insights into the variability of drug usage, adherence, and prescribing patterns.

Average gap rate for N06 drugs

Drug Name	Average Gap Rate
Amitriptyline	0.499649
Nortriptyline	0.469104
Trimipramine	0.453113
Trazodone	0.435864
Doxepin	0.401310
Imipramine	0.391257
Dothiepin	0.373427
Mianserin	0.369716
Clomipramine	0.368116
Amoxapine	0.351190
Maprotiline	0.324706
Protriptyline	0.265053
Phenelzine	0.257697
Nefazodone	0.240160
Mirtazapine	0.208985
Desipramine	0.191415
Duloxetine	0.186261
Fluvoxamine	0.178815
Venlafaxine	0.174869
Reboxetine	0.172705
Citalopram	0.165329
Bupropion	0.142018
Escitalopram	0.138218
Paroxetine	0.135748
Moclobemide	0.133249
Lofepramine	0.127883
St. John's Wort Extract	0.126383
Fluoxetine	0.117253
Tryptophan	0.109548
Sertraline	0.095081
Tranylcypromine	0.090527
Vortioxetine	0.052356

Table 15: Average gap rates for N06 drugs. Higher gap rates indicate more frequent interruptions in drug usage.