Medicine Prescription Pattern Analyzer for Drug Utilization

Arpit Yadav
School of Computer Science and Engineering
Lovely Professional University, Punjab, India
arpit.yadav232@lpu.in

Harsh Tripathi
School of Computer Science and Engineering
Lovely Professional University, Punjab, India
harshtripathi5954@gmail.com

***Abstract*—Drug Utilization Management (DUM) sits at the center of healthcare systems we build to keep drug use sensible, affordable, and steady over time. Places like India, where there's a flood of brand-name drugs hitting the market, see medication choices shaped not just by what the clinical guidelines say, but also by things like pricing gaps, a few big manufacturers steering the scene, and whether a drug is even available to begin with.**

The usual approach to studying drug use leans on data from individual doctors or patients - information that's almost always locked away because of privacy rules, regulations, or tight-fisted institutions. So, we put together a market-driven system for analyzing drug use risks and supporting prescription decisions, built with unsupervised machine learning in mind.

We dug into data from more than 250,000 branded drugs sold in Indian pharmacies. Before we got into the modeling, we spent a lot of time cleaning and shaping the data to figure out which factors really capture price movement, how crowded the market is, and whether supply stays steady. Since there's no labeled data showing actual drug usage results, we turned to Isolation Forest for finding outliers, and used that to estimate a running risk score for different chemicals.

The model we designed can spot drug combos that are overpriced, have barely any competition, and are more likely to vanish from shelves. On top of that, we built a simple rule-based tool to help with prescriptions - giving cheaper and available alternatives within the same chemical group. Using unsupervised learning on this kind of market data means our system offers healthcare providers a private, straightforward way to flag risky drug trends and suggest better options, especially in places where detailed patient data is scarce.

***Index Terms*—Drug Utilization Management, Pharmaceutical Market Analysis, Medicine Price Variation, Anomaly Detection, Isolation Forest, Prescription Support, Healthcare Analytics**

I. Introduction

Drug Utilization Management (DUM) is more than a simple checklist to ensure the right drug is prescribed at the right dose for the right duration. It is a broader effort to make sure medicines are used sensibly, remain affordable, and are not wasted. This becomes especially important in settings where patients pay for healthcare out of their own pockets and where inefficient or irrational drug use drives up costs, putting additional pressure on hospitals and clinics that are already operating at their limits.

In real-world clinics and pharmacies, however, drug use is shaped by much more than clinical guidelines or what textbooks recommend. While protocols define what is considered best practice, doctors, procurement teams, and patients also have to weigh practical concerns. These include how much a medicine costs, which brands are actually available, the reputation of the manufacturer, marketing pressure, and whether the supply is reliable. Such factors play an even bigger role in markets dominated by branded generics, where different brands contain the same active ingredient but are sold at vastly different prices.

India is a clear example of how complex this situation can become. Its pharmaceutical market is massive and highly fragmented. For a single drug molecule at a given strength, there may be dozens of brands sitting side by side on pharmacy shelves. In most cases, the tablets are chemically identical, yet their prices vary dramatically. Numerous studies have shown that one brand can cost several times more than another without offering any added therapeutic benefit.

For patients, choosing a higher-priced brand can quickly translate into financial strain. In some cases, the burden becomes so heavy that patients begin skipping doses or abandon treatment altogether. From a system-wide perspective, these choices waste limited healthcare resources and reduce access to essential medicines. When families are paying out of pocket, the consequences are even more severe.

Traditional drug utilization studies have helped track prescribing patterns and identify problem areas over time. Most rely on prescription records or patient data and compare them against benchmarks set by the World Health Organization (WHO). While valuable, this approach has clear limitations. Accessing such data is often difficult due to privacy concerns, ethical approvals, and administrative barriers. As a result, real-time, large-scale monitoring of drug use remains challenging.

Market-level data offers a different perspective, closer to a bird's-eye view of the system. These datasets capture information on pricing, manufacturers, formulation details, and product availability across regions. Although they cannot directly measure clinical outcomes, they are useful for identifying unusual price differences, market dominance by a few players, or weaknesses in the supply chain. All of these factors quietly influence prescribing decisions every day.

As pharmaceutical market data continues to grow, and as machine learning tools become more accessible, new opportunities are emerging for drug utilization management. The challenge is that, unlike controlled clinical datasets, market data does not come with clear labels indicating what is "good" or “bad.” This makes traditional supervised learning methods less suitable. Instead, techniques such as unsupervised learning and anomaly detection are better suited to uncover patterns, identify outliers, and flag drugs that may warrant closer scrutiny due to unusual or potentially risky behavior.

This study presents a market-driven framework for analyzing drug utilization risks and supporting prescription decisions using unsupervised machine learning applied to large-scale pharmaceutical market data. Rather than relying on patient records or prescription logs, the approach focuses on drug composition and market-related attributes. High-risk medicines are identified using only market-level signals. The key contributions include: (i) developing composition-level features that reflect pricing behavior, market competition, and supply stability; (ii) applying an Isolation Forest–based anomaly detection model to generate a continuous utilization risk score; and (iii) proposing a prescription support mechanism grounded entirely in market data.

II. Literature Review

A. Drug Utilization Studies and Rational Use of Medicines

Drug utilization research is a well-established field that looks at how medicines are used across populations, with the broader goal of encouraging rational therapy, improving patient outcomes, and reducing avoidable healthcare costs. At its core, rational drug use means that patients receive medicines that genuinely match their clinical needs, in doses tailored to the individual, for an appropriate length of time, and at the lowest possible cost to both the patient and the wider community.

The World Health Organization (WHO) helped formalize this area of research by introducing a set of standardized prescribing measures, commonly known as the WHO/INRUD indicators. These include indicators such as the average number of medicines per prescription, the extent of generic prescribing, the share of medicines drawn from essential medicines lists, and the frequency of antibiotic and injection use [1]. Over time, these indicators have become a standard tool in hospital- and community-based studies, allowing researchers to assess prescribing behavior and flag patterns of irrational medicine use.

In India, a substantial body of research has applied these WHO prescribing indicators to study drug utilization across different tiers of the healthcare system. Findings from systematic reviews and individual studies repeatedly point to familiar concerns, including polypharmacy, a strong preference for branded medicines over generics, and wide variation in prescribing practices between institutions [2]. These issues are particularly pronounced in outpatient and primary care settings, where cost considerations often play a decisive role in treatment choices.

More recent work has drawn attention to the ongoing disconnect between policy-level recommendations and what happens in everyday clinical practice. For example, Goruntla et al. [3] showed that even though essential medicines lists are readily available, adherence to rational prescribing guidelines remains uneven. Along similar lines, Joshi et al. [4] reported incomplete prescriptions and frequent departures from recommended practices in tertiary care hospitals, highlighting the limits of relying on guidelines alone to drive change.

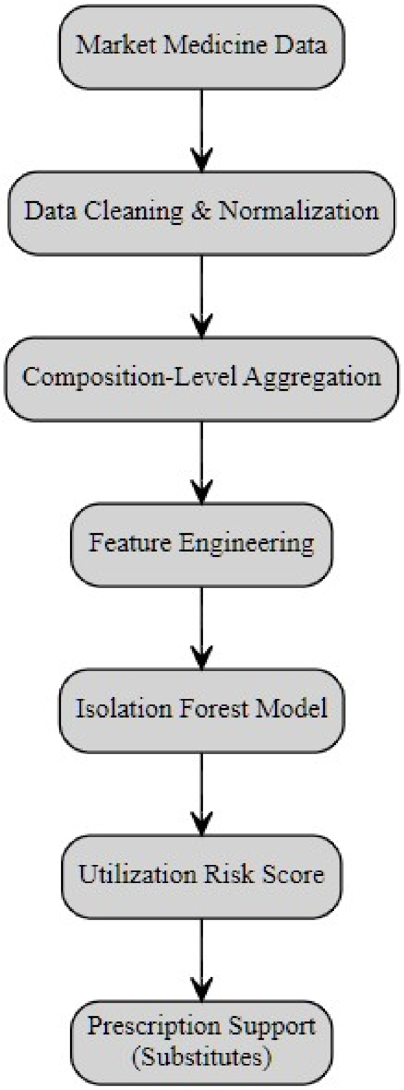
Market Medicine Data
Data Cleaning & Normalization
Composition-Level Aggregation
Feature Engineering
Isolation Forest Model
Utilization Risk Score
Prescription Support (Substitutes)

Fig. 1. Overall workflow of the proposed market-driven drug utilization risk analysis framework, illustrating data preprocessing, composition-level feature engineering, unsupervised risk modeling using Isolation Forest, and prescription support through substitute identification.

Despite their clear clinical value, traditional drug utilization studies are not without limitations. Most depend on manually collected prescription samples, which tend to be small, time-bound, and not always representative of broader practice patterns. Access to prescription-level data is also often restricted by privacy regulations, ethical approval processes, and institutional data governance rules. Together, these challenges make it difficult to scale prescription-based monitoring efforts or sustain them over time, especially in healthcare systems with limited resources.

B. Medicine Price Variation and Market Competition

Medicine price variation plays a major role in how drugs are actually used, especially in markets dominated by branded generics. In India, it is common to find multiple brands selling medicines with the same chemical composition and strength, yet priced very differently. This coexistence of therapeutically equivalent products has led to wide and often striking price differences across the market.

A growing body of research has documented just how extreme these price gaps can be for commonly prescribed medicines in India. Ray [5], for instance, carried out a detailed cost variation analysis and found price differences running into several hundred percent between the cheapest and the most expensive brands of the same formulation. For patients, this is not just a numbers problem. Higher prices can quickly make treatment unaffordable and may push economically vulnerable patients to skip doses or stop therapy altogether.

These price differences are rarely explained by manufacturing costs alone. Instead, they are shaped by the structure of the market itself, branding and marketing strategies, and gaps in information. Both prescribers and patients often assume that a higher price signals better quality, even when there is no clinical evidence to support that belief. When such perceptions guide prescribing decisions, rational drug use suffers, and policies aimed at price regulation or promoting generics lose much of their intended impact.

Competition within the pharmaceutical market also has a direct bearing on pricing behavior and the reliability of supply. Drugs produced by only a handful of manufacturers are more exposed to monopolistic pricing and are at greater risk of supply disruptions. Research on market dynamics consistently shows that lower levels of competition tend to drive prices upward and make shortages more likely, particularly when dominant manufacturers withdraw certain products or formulations.

Taken together, these observations point to the value of including market-level indicators, such as price distribution, brand density, and manufacturer participation, in drug utilization studies. Unlike analyses based solely on prescriptions, market-based approaches can reveal broader, system-level risks related to affordability and supply continuity, both of which have a direct and lasting influence on how medicines are used in everyday practice.

C. Machine Learning in Healthcare and Pharmaceutical Analytics

The explosion of healthcare and pharmaceutical data has pushed machine learning to the forefront of data-driven decision support. Today, these techniques are used for a wide range of tasks, from predicting disease risk and analyzing treatment outcomes to spotting fraud and making better use of healthcare resources.

In practice, though, many healthcare datasets do not come with clear or reliable labels, especially for questions tied to utilization patterns or costs. Terms like "irrational prescribing" or "unsafe use" are inherently subjective and usually require expert clinical judgment, along with access to detailed patient-level information. Because of these constraints, supervised learning approaches are often difficult to apply in real-world settings.

This is where unsupervised learning methods come into play. By design, they look for structure and patterns in data that have no predefined labels. Anomaly detection, in particular, is well suited to highlighting observations that stand out from typical behavior and may signal elevated risk. In healthcare analytics, such methods have already been used to flag unusual patient activity, detect fraudulent insurance claims, and identify unexpected clinical events.

Isolation Forest is one of the most commonly used algorithms for unsupervised anomaly detection. It works by randomly partitioning the feature space, which makes it easier to isolate data points that behave differently from the norm [6]. Unlike distance- or density-based approaches, it does not depend on strict distributional assumptions, allowing it to scale efficiently to high-dimensional datasets. This combination of efficiency and interpretability has made it popular across domains such as healthcare, finance, and industrial monitoring.

When applied to pharmaceutical analytics, anomaly detection provides a practical way to identify medicines with unusual pricing patterns, limited market competition, or unstable supply. Viewing drug utilization risk through the lens of anomaly detection makes it possible to carry out large-scale risk assessments without relying on labeled outcomes or rigid thresholds, offering a flexible and scalable solution.

D. Comparative Summary of Related Work

Table I summarizes representative studies related to drug utilization research, medicine price variation, and machine learning applications in healthcare. The comparison highlights differences in data sources, analytical approaches, and key limitations.

E. Research Gap

A review of the literature shows that drug utilization research, medicine price variation studies, and machine learning applications in healthcare have mostly developed along separate tracks. Traditional utilization studies tend to focus on clinical prescribing behavior, but they are often limited by restricted access to prescription data and relatively small sample sizes. On the other hand, market-based price analyses are effective at highlighting economic inefficiencies, yet they usually stop short of offering integrated frameworks for assessing utilization risk.

At the same time, most machine learning work in healthcare has centered on supervised clinical prediction tasks. There has been far less attention on using unsupervised methods to assess utilization risk, particularly when relying on market-level data. Only a handful of studies have explored the potential of combining large-scale pharmaceutical market datasets with anomaly detection techniques to support drug utilization management.

TABLE I
COMPARISON OF RELATED WORK IN DRUG UTILIZATION AND PHARMACEUTICAL ANALYTICS

Study	Data Source	Methodology	Focus Area	Key Limitations
Shalini et al. [1]	Prescription samples	WHO indicators	Rational drug use assessment	Small sample size, manual data collection
Gujar et al. [2]	Literature review	Systematic review	Prescribing patterns in India	Lack of quantitative risk modeling
Ray [5]	Market price data	Cost variation analysis	Brand-level price dispersion	No utilization risk scoring
Joshi et al. [4]	Hospital prescriptions	Indicator-based analysis	Prescribing completeness	Limited scalability
Liu et al. [6]	Synthetic/benchmark data	Isolation Forest	Anomaly detection	Not applied to pharmaceutical markets

This study aims to bridge these gaps by introducing a unified, market-driven framework that integrates pharmaceutical market data with unsupervised machine learning to quantify utilization risk at the level of chemical composition. The proposed approach is designed to be scalable, interpretable, and respectful of data privacy, making it well suited for use in healthcare settings where access to detailed patient data is limited.

III. DATASET DESCRIPTION AND EXPLORATORY ANALYSIS

A. Dataset Source and Scope

The dataset used in this study provides a detailed snapshot of the Indian pharmaceutical market and was compiled from publicly available medicine information sources. It includes 253,973 records, with each entry representing a branded medicine available in India as of November 2022. The dataset brings together products from several thousand pharmaceutical manufacturers, spanning large multinational companies as well as small and medium-sized domestic firms.

Each record contains a combination of pharmaceutical and commercial details, such as the brand name, chemical composition, formulation type, dosage strength, pack size, manufacturer, listed price, and current availability. A dedicated indicator flags medicines that have been discontinued, which makes it possible to examine supply continuity and overall market stability. Unlike prescription-based datasets, this collection does not include any patient-, prescriber-, or outcome-level information, ensuring adherence to data privacy and ethical research standards.

Because it captures pricing behavior and supply-side dynamics, the dataset is well suited for market-driven analyses of drug utilization. Its scale and diversity support robust statistical analysis across a broad range of therapeutic areas and market segments.

B. Data Cleaning and Preprocessing

Raw pharmaceutical market data often come with their fair share of issues, including inconsistencies, duplicate entries, and mixed formatting caused by differences in data sources and reporting standards. To ensure the analysis rested on solid ground, the data went through extensive cleaning and preprocessing before any exploration or modeling was carried out.

Text-based fields such as medicine names, manufacturer names, and chemical compositions were standardized by normalizing letter case and stripping out unnecessary characters. Duplicate entries, often created by repeated listings of the same brand, were identified and removed to avoid skewing the results. Price information was converted into a consistent numeric format, and records with missing, zero, or clearly invalid prices were excluded from further analysis.

The availability status was then simplified into a binary indicator that distinguished between medicines currently on the market and those that had been discontinued. This made it easier to analyze discontinuation trends across different drug compositions and manufacturers. Medicines containing multiple active ingredients were flagged and handled separately to ensure that composition-level characteristics were represented accurately.

C. Composition-Level Aggregation

A key design decision in this study was the shift from analyzing individual brands to focusing on chemical compositions. While brand-level analysis reflects commercial positioning and marketing differences, a composition-level view aligns more closely with therapeutic equivalence and clinical interchangeability.

To support this shift, brand-level records were aggregated for each unique chemical composition to produce a set of composition-level indicators. These included summary measures such as average and maximum price, the number of brands available, the number of distinct manufacturers, and the share of brands that had been discontinued. Aggregating the data in this way helped cut through the noise created by brand proliferation and allowed for more meaningful comparisons among medicines that are therapeutically equivalent.

This composition-based perspective is especially relevant in the Indian pharmaceutical market, where many brands selling the same formulation coexist, often with wide price differences and uneven manufacturer participation. By concentrating on chemical compositions rather than individual brands, the analysis highlights factors that matter for drug utilization and access, rather than differences driven largely by branding and promotion.

D. Dataset Attribute Summary

Table II summarizes the key attributes included in the dataset and their role in the analysis. The attributes were selected to capture pricing behavior, market structure, and availability dynamics relevant to drug utilization management.

E. Exploratory Analysis of Pricing Behavior

Exploratory analysis showed clear differences in medicine prices across chemical compositions. Overall, the price distribution was strongly right-skewed, with a small number of medicines priced very high and a much larger group falling into the mid-price range.

Looking more closely at the composition level, prices often varied sharply even when the dosage form and strength were exactly the same. In several cases, the most expensive brand for a given composition cost several times more than the average, pointing to the coexistence of premium brands alongside much cheaper alternatives.

This level of price variation within a single composition has real consequences for how medicines are used. When doctors or patients opt for higher-priced brands despite the availability of equally effective, lower-cost options, affordability suffers and avoidable inefficiencies creep into the healthcare system. Taken together, these findings highlight the importance of accounting for both typical prices and extreme price points when assessing utilization risk.

F. Market Competition and Manufacturer Participation

An analysis of brand and manufacturer counts showed wide differences in how competitive the market is across chemical compositions. Many commonly prescribed compositions were linked to a large number of brands and manufacturers, pointing to a healthy level of competition. These medicines typically showed more moderate pricing and were less vulnerable to supply disruptions.

By contrast, some compositions were produced by only a handful of manufacturers, leading to much more concentrated markets. In these cases, prices tended to be higher and supplies less stable, especially when dominant manufacturers withdrew particular brands or formulations. Manufacturer diversity emerged as a key factor shaping both pricing behavior and product availability.

Taken together, these findings underscore the role of market competition in moderating drug utilization risk. Compositions supported by a broader base of manufacturers are more likely to have stable prices and resilient supply chains, which, in turn, supports more rational use of medicines..

G. Availability and Discontinuation Patterns

Availability analysis showed that although most medicines in the dataset were still actively marketed, a noticeable share had already been discontinued. These discontinuation trends were not uniform and differed widely depending on the chemical composition and the manufacturers involved. Medicines linked to a larger number of brands tended to have lower discontinuation rates, pointing to a more stable market presence. In contrast, compositions produced by only a few manufacturers were more prone to discontinuation, which can disrupt treatment continuity and force a shift to alternative therapies. The existence of discontinued brands within a given composition offers useful insight into underlying supply-chain behavior and long-term market sustainability. Factoring discontinuation-related indicators into drug utilization analysis therefore helps flag compositions that may carry a higher

H. Implications for Feature Design

The exploratory analysis revealed the features relevant to the utilization that were subsequently employed in modeling. The pricing behavior, market competition, and availability came out as the main forces that determined the drug utilization in the Indian pharmaceutical market. The dataset was changed into a structured representation that was fit for the unsupervised machine learning-based risk analysis by converting these dimensions into quantitative features such as average price, maximum price, number of brands, number of manufacturers, and discontinuation ratio. These features constitute the basis for the formulation of the problem and the modal approach which is described in the next section.

IV. Problem Formulation and Feature Engineering

A. Problem Definition

This research aims to uncover the composition of drugs that are in the highest risk group for being utilized because of bad market conditions. This study diverges from traditional drug utilization studies, which mainly use prescription-level data to focus on clinical appropriateness, by treating drug utilization risk as a market-driven analytical problem.

We define $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ to be the collection of distinct chemical mixtures extracted from drug market data. For every mixture c_i, a number of attributes derived from the market about the mixing behavior, market competition, and supply chain are assigned to it. The task at hand is to measure a relative utilization risk score $R(c_i)$ for every mixture without the provision of labeled utilization results.

In this respect, utilization risk indicates the chance that the employment of a given mixture may cause economic inefficiency, decreased accessibility, or supply instability in case of routine health care practice deployment. The problem is already unsupervised because there are no ground-truth labels indicating ”high-risk” or ”low-risk” compositions available.

B. Rationale for Unsupervised Learning

In actual pharmaceutical analytics, there is a lack of labeled datasets fit for supervised learning purposes. The labels like irrational prescription, inappropriate utilization, or unsafe medicine use are determined by medical experts and require access to comprehensive patient-level data. The process of upscaling such labeling is not only resource-demanding but also ethically restricted.

On the other hand, unsupervised learning is a practical alternative since it allows the detection of strange or abnormal patterns without any prior labels. The learning process of the unsupervised models is based on typical market behavior so that the researchers can spot the market segments that differ

TABLE II
SUMMARY OF DATASET ATTRIBUTES AND ANALYTICAL ROLE

Attribute	Type	Description	Analytical Role
Medicine Name	Text	Brand name of the marketed medicine	Brand identification
Chemical Composition	Text	Active ingredient(s) and strength	Composition-level aggregation
Formulation Type	Categorical	Dosage form (tablet, capsule, injection, etc.)	Therapeutic equivalence
Manufacturer Name	Text	Producing pharmaceutical company	Market competition analysis
Price	Numeric	Listed market price of the medicine	Economic burden assessment
Availability Status	Binary	Available or discontinued	Supply stability analysis
Pack Size	Numeric	Units per package	Price normalization

greatly from the standard pricing, competition, and commodity flow patterns. Such deviations then act as signals for a possible risk of improper utilization.

This approach is in line with the support for decision-making characteristic of drug utilization management where the aim is not to substitute medical judgment but rather to direct medicines for further checking and possibly, being intervened.

C. Composition-Level Feature Representation

The feature engineering process took place considering chemical composition as the main criterion so that the analysis would hinge on therapeutic equivalence rather than brand-level commercial differentiation. A feature vector $\mathbf{x}_i \in \mathbb{R}^d$ was formed for each composition c_i, with the corresponding dimension assigned to a market indicator that is relevant to use.

Let $\mathcal{B}_i$ be the notation for the brands that are connected with composition c_i. At this point, brand-level characteristics were combined to create composition-level features by means of summary statistics like the average, maximum, and proportions. The process of aggregation not only removes the noise associated with brand proliferation but also makes the results easier to understand.

D. Pricing-Based Features

Pricing behavior is a primary determinant of drug utilization, particularly in healthcare systems with high out-of-pocket expenditure. To capture pricing-related risk, multiple price-based features were engineered.

The average price of a composition is defined as:

$$\mu_i = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} p_b$$

where p_b denotes the listed price of brand b.

The maximum price captures extreme pricing behavior:

$$p_i^{\max} = \max_{b \in \mathcal{B}_i} p_b$$

The factors mentioned together give the typical and the worst-case economic burden of a composition. Logarithmic transformations were applied to the price features before the modeling process to reduce skewness and stabilize variance.

E. Market Competition Features

Pricing and supply resilience are both directly affected by market competition. It is a common understanding that compositions with a greater number of plant-based brands and manufacturers tend to be less susceptible to price monopolies and supply interruptions.

Two competition-related features were engineered:

$$n_i^{\text{brand}} = |\mathcal{B}_i|$$

$$n_i^{\text{manuf}} = |\mathcal{M}_i|$$

where $\mathcal{M}_i$ denotes the set of manufacturers producing brands associated with composition c_i.

Higher values of n_i^{brand} and n_i^{manuf} indicate a competitive market structure, while lower values suggest concentration and potential utilization risk.

F. Availability and Discontinuation Features

Supply continuity is a critical component of rational drug utilization. Medicines with unstable availability or high discontinuation rates can disrupt treatment regimens and force substitution with potentially more expensive alternatives.

To quantify supply-related risk, a discontinuation ratio was computed for each composition:

$$d_i = \frac{|\mathcal{B}_i^{\text{disc}}|}{|\mathcal{B}_i|}$$

where $\mathcal{B}_i^{\text{disc}}$ denotes the subset of discontinued brands associated with composition c_i.

A higher discontinuation ratio indicates greater supply instability and elevated utilization risk.

G. Relative Overpricing Indicator

In addition to absolute pricing measures, a relative overpricing indicator was derived to capture deviations from composition-level pricing norms. For each brand $b \in \mathcal{B}_i$, the relative price deviation was computed as:

$$\delta_b = \frac{p_b - \mu_i}{\mu_i}$$

These deviations were aggregated to derive a composition-level overpricing indicator:

$$o_i = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} \max(\delta_b, 0)$$

This feature highlights compositions where a substantial proportion of brands are priced significantly above the average, suggesting potential information asymmetry or irrational pricing practices.

H. Feature Normalization and Interpretability

The model features were first standardized to zero mean and unit variance, thus making the dimensions comparable. Logarithmic transformations on the price-related features decreased the outlier's impact but the relative differences were still preserved.

Every feature created has a clear interpretation linked to the dynamics of the pharmaceutical market which makes analysis and trust of the stakeholders transparent. The final feature set is a compact but expressive representation of the utilization-relevant characteristics forming the basis of the unsupervised learning methodology described in the next section.

V. Machine Learning Methodology

A. Modeling Objective

The machine learning part's aim is to find out drug combinations that show unusual market behavior that signals the risk of utilization to be increased. Instead of predicting a predetermined clinical or economic outcome, the model gives each chemical composition a relative risk score based on how much they differ from the typical patterns of pricing, competition, and availability observed in the pharmaceutical market.

Since there are no labeled utilization outcomes, the modeling task is exploratory by nature and prioritization-oriented as well. The aim is not to categorize medicines as good or bad but to order compositions based on their relative utilization risk and thus facilitating the healthcare stakeholders' targeted review and intervention.

B. Choice of Modeling Paradigm

The supervised learning methods necessitate labeled data that categorically specify the link of a drug with either irrational use or adverse effects. In the pharmaceutical market analytics, such labels are very rarely present in large numbers and normally they combine the expert clinical judgement with the very detailed patient-level data to be formed. Hence, the supervised classification methods cannot be applied to today's problem.

The unsupervised learning methods, which directly infer structure from unlabeled data, are more compliant to the limitations of market-based utilization analysis. By learning the distribution of the typical market behavior, the unsupervised models can detect the compositions that differ significantly from the normative patterns. These deviations are then interpreted as signs of possible utilization risks.

C. Rationale for Isolation Forest

The Isolation Forest technique was chosen among unsupervised anomaly detection methods due to its ability to scale, be robust, and suitable for high-dimensional data. Unlike distance-based methods like k-means clustering or density-based methods like DBSCAN, Isolation Forest does not make explicit assumptions concerning the form of the data or the structure of the clusters.

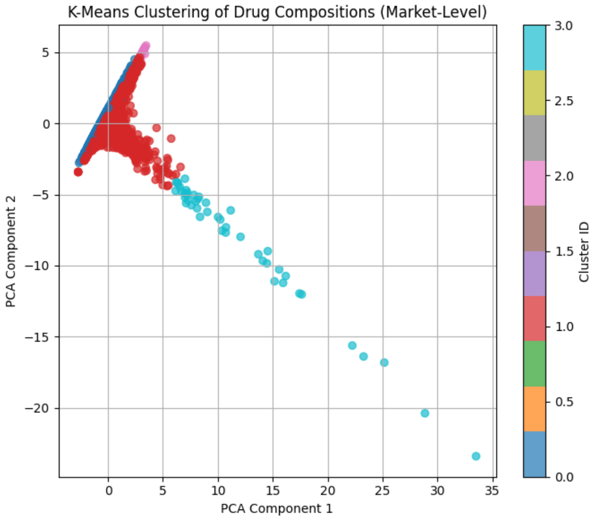

Fig. 2. K-Means Clustering of Drug Compositions

The Isolation Forest algorithm works by continuously cutting the feature space with the help of randomly chosen features and split values. If fewer splits were used to separate a point from the rest, that point is regarded as being more anomalous. This approach is especially good for discovering rare and exceptional behaviors in large datasets.

For the formal definition, the anomaly score for the sample c_i is computed based on the average path length used to separate its feature vector $\mathbf{x}_i$ over the ensemble of isolation trees. More anomalous samples have shorter average path lengths.

D. Comparison with Alternative Unsupervised Methods

Different unsupervised methods were investigated but ultimately considered unsuitable for the current situation. One such method was the k-means algorithm which is a clustering technique requiring the number of clusters to be defined beforehand and relying on the existence of spherical cluster shape which is an incorrect assumption for the somewhat chaotic and mixed nature of pharmaceutical market data.

Similarly, density-based methods like LOF (Local Outlier Factor) and DBSCAN are heavily dependent on the scaling of features and parameter setting and may not be very effective in the market data that has various density areas. One-class SVMs (support vector machines) carry a theoretical foundation for the anomaly detection but their scalability to large datasets is really poor along with the requirement of careful kernel selection.

However, the Isolation Forest has the linear number of samples scale, excellent power in dealing with mixed feature distributions and finally provides a continuous anomaly score, which is very appropriate for decision support through ranking.

These very factors were behind the selection of this method as the core modeling approach.

E. Model Input and Preprocessing

The Isolation Forest technique utilizes the feature vectors at the composition level as its input, which are derived from the feature engineering stage. In advance of the training, a

standardization step was done to all features so that they would all be on the same scale and to prevent variables having a large range of values from dominating the others.

The features connected to the price were logarithmically transformed which stood to lower their right skewness and make the variance constant. As a result of this transformation, very high-priced compositions will not drive the partitioning process while the relative price differences will still be there.

F. Model Configuration and Training

The Isolation Forest model was trained by means of a batch of decision trees created through random feature selection and random split threshold combinations. A large number of trees were used to ensure that the estimation of anomaly scores was stable throughout the different compositions.

The contamination rate i.e. the percentage of potentially contaminated observations was set at a low level to focus on extreme cases of utilization risk only. This choice is in line with the decision-support nature of the system, where false positives could lead to unnecessary scrutiny.

Random seed initialization was applied to ensure that the results would be the same every time. During model training, there is no iterative optimization or gradient-based learning, resulting in efficient training even for large datasets.

G. Utilization Risk Score Definition

The unprocessed anomaly score given by the Isolation Forest model is an indication of how much each composition diverges from the normal market behavior. In order to make it easier to interpret, this score was converted into a utilization risk score $R(c_i)$ that is normalized to the range $[0, 1]$, with higher values meaning higher relative risk.

The utilization risk score indicates a continuous ranking of compositions instead of a binary classification. The representation based on ranking permits the stakeholders to set flexible limits according to the operational restrictions and policy goals.

H. Interpretation of Model Output

Compositions with high utilization risk scores usually display one or more of the following traits: very high average or maximum prices, small brand and manufacturer involvement, high discontinuance rates, or strong overpricing compared to composition-level standards.

However, a high utilization risk score should not be taken as a negative sign for the drug in question or that it is no longer doctor-prescribed. Rather, it points to possible economic inefficiency or vulnerability regarding supply that might be wanting a review. This difference makes sure that model predictions are understood within a proper decision-making aid system.

I. Computational Complexity and Scalability

Isolation Forest operates with a linear time complexity concerning both the number of samples and features thus, it is appropriate for very large datasets from the pharmaceutical market. The memory demand is not too high since the isolation trees only keep split rules instead of complete distance matrices.

The suggested framework's computational efficiency allows its practical application in actual healthcare analytics scenarios, such as integration with dashboards, procurement systems, and prescription support tools.

VI. Results and Case Studies

A. Overview of Model Output

The Isolation Forest model, which had been trained, assigned each chemical composition a risk score in terms of utilization. The scores are a reflection of the different degrees of the pharmaceuticals that are composed of the particular composition deviating from the normal market behavior in relation to price, competition, and availability. The model does not result in a binary classification, but it does allow for the ranking of compositions according to the level of risk from the lowest to the highest.

The analysis of the score distribution showed a very skewed pattern; most of the compositions were in the low to moderate risk group while a smaller group had very high-risk scores. This splitting up of the groups implies that only a small number of the compositions are displaying the market characteristics so severe that they need to be closely watched.

B. Identification of High-Risk Compositions

The compositions that were given high scores of utilization risk showed consistently one or more unfavorable market characteristics. Some of those were extremely high average or maximum prices, a very small number of available brands, low manufacturer diversity, and high discontinuation ratios. It is important to note that high utilization risk was hardly ever a result of a single factor; it was more of a cumulative effect of several negative indicators.

Table III gives a representative example of high-risk compositions detected by the model together with the primary contributing features. The composition names have been kept secret to highlight the relevance of the methodology rather than the specific commercial products.

C. Pricing Behavior Analysis

An extensive scrutiny of features linked to pricing revealed that a number of high-risk materials were linked to very high price dispersion. In a few instances, a handful of brands that were priced at the top of the market level were the ones that pushed the maximum price metric up while they were at the same time cheaper than the alternatives. These trends are an indication of possible inefficiencies that could result from, for instance, information asymmetry or brand-oriented doctor prescribing practices.

TABLE III
REPRESENTATIVE HIGH-RISK DRUG COMPOSITIONS IDENTIFIED BY THE MODEL

Composition ID	Avg. Price	Max Price	No. of Brands	Discontinuation Ratio
C_A	High	Very High	Low	High
C_B	Moderate	High	Very Low	Moderate
C_C	High	High	Low	Low
C_D	Moderate	Very High	Moderate	High

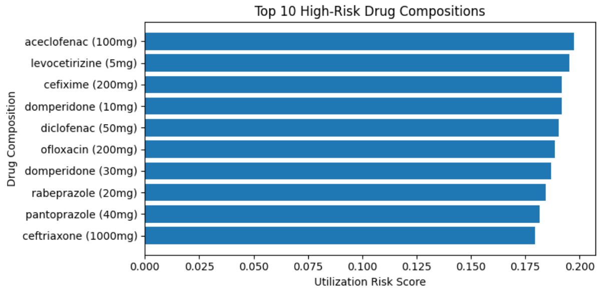

Fig. 3. K-Means Clustering of Drug Compositions

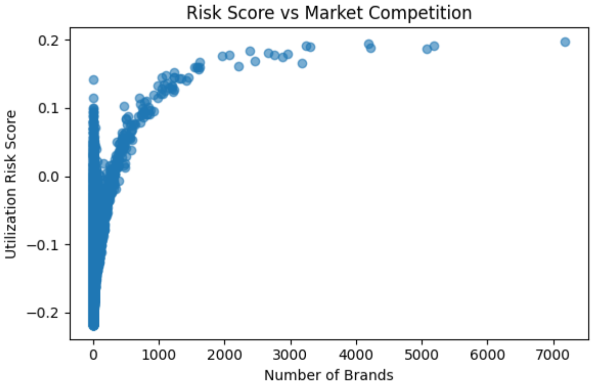

Fig. 4. K-Means Clustering of Drug Compositions

The model, however, did not automatically categorize all high-priced compositions as high-risk ones. Some specialized or life-saving drugs that were priced uniformly high across brands were allotted moderate-risk scores owing to stable market structure and consistent manufacturer participation. This difference highlights the model's power to separate between the forces of nature that impose expensive treatments and those of the market that drive up prices of medicines irrationally.

D. Market Competition and Manufacturer Diversity

Market competition came out as the strongest moderating factor in utilization risk. Compositions characterized by a larger number of competing brands and manufacturers usually had lower risk scores assigned to them even if the prices on the average were comparatively high. The existence of several suppliers minimizes the risk of monopoly pricing and also supports the supply chain in times of high demand.

In contrast, compositions that were controlled by a few manufacturers were more often conducted as high-risk ones. The limited competition situation not only raises the affordability of the price hikes but also it makes the area more prone to the disruptions of supply. The mentioned phenomena underline the significance of manufacturer diversity as a fundamental element of drug utilization rationality.

E. Availability and Discontinuation Effects

Availability-related aspects were vital in establishing utilization risk. The compounds with a great discontinuation ratio were at risk of receiving high-risk scores more especially together with few remaining suppliers. Such combinations are a lot harder for long-term treatment and may cause doctors to change the patients' medications to other, usually more expensive, ones due to the high cost.

On the other hand, the mixtures with low discontinuation ratios and continuous gut supply across multiple brands were seen to have less utilization risk even when the price was moderately varying. This experience points out that supply stability is a primary factor of utilization risk.

F. Case-Based Interpretation of Utilization Risk

In order to make a practical application of the model output, several composition-level scenarios were qualitatively examined. In one typical case, a very prescribed composition showed a big gap in brand prices with a handful of brands having much higher prices than the average for the composition. Even though there were plenty of alternatives that were affordable, the presence of the overpriced brands still managed to increase the overall risk score, indicating that prescribing behavior was potentially economically inefficient.

Another example is of a composition that had a few brands and manufacturers but still had a moderate average price accompanied by a high discontinuation ratio. No extreme pricing was present, yet the heightened risk score was pointing to possible supply vulnerability instead of direct economic burden. Such insights are hard to come by in prescription-level studies but are easily traced via market-based analysis.

G. Prescription Support Through Substitute Identification

The suggested framework does not only focus on risk recognition; it notably features a rule-based support mechanism for prescriptions intended to mitigate the risk of over-utilization. In case of a mixture being classified as high-risk, the system not only identifies other stocks of the same composition but also checks if their prices are lower than the average of the composition price.

An example of the substitution identification is shown in Table IV for a composition with high-risk. The system pushes the selection of non-therapeutically equivalent medicines by uncovering the most cost-effective and available alternatives.

H. Summary of Key Findings

The research shows that the unsupervised machine learning algorithm used for pharmaceutical market data could successfully pinpoint the risks related to utilization that were caused by the pricing practices, market density and instability of supply. The high utilization risk is not the result of one single market factor or an isolated anomaly; rather it comes from the interplay of several market factors.

The integration of risk scoring and prescription support not only helps to identify the problem but also provides solutions for health care managers, purchasing departments and doctors. The outcome of the research supports the framework that was put forward as an effective instrument for controlling drug utilization based on market factors.

VII. Discussion, Limitations, and Policy Implications

A. Interpretation of Findings

The research that was carried out has concluded that if market-driven data related to pharmaceuticals is used properly along with unsupervised machine learning, it can actually lead to apprehension about drug consumption risk. This innovative technique, which considers the problem of utilization risk as anomaly detection, reveals the complex relationships between pricing policies, market competition, and supply issues that are difficult to identify through conventional analytics.

One big takeaway is that high-risk utilization is seldom tackled by one adverse factor alone. On the contrary, risky combinations usually consist of overpricing, lack of competition among manufacturers, and erratic supply. This insight reinforces drug utilization management by multidimensional analysis, as relying on any one indicator, for instance only price, could lead to incomplete or distorted conclusions.

Nonetheless, signaling of high drug usage risk should not be seen as an evaluation of the appropriateness or the value of the therapy clinically. A number of the mixtures that were flagged as high-risk could be of great importance or even life-saving from a clinical standpoint but at the same time could be creating trouble economically or in supply. The framework is, therefore, designed to be an aid for making informed decisions rather than a replacement for clinical judgment.

B. Comparison with Conventional Drug Utilization Approaches

The conventional drug utili zation studies have been primarily concerned with the analysis of prescribing behavior through either prescription-level or patient-level data. The prescriptive studies have been offering valuable inputs to the medical decision making process; however, they are usually tainted with the issues of limited data access, small sample sizes as well as manual data collection processes.

The suggested market-driven approach not only competes with such studies, but it also complements the traditional ones by reallocating the analytical focus from patient-level data to entire compositions of drugs sold in the market. The redirection of the study areas makes it possible to perform large-scare analyses without the need for sensitive clinical data thus making the approach both scalable and privacy-preserving. Furthermore, the application of unsupervised learning reduces the burden of dependence on subjective or institutional-specific labeling criteria thus increasing generalizability across healthcare settings.

The market-driven framework should not be seen as a competitor to prescription-based studies, but rather as a tool that adds another layer of analysis from the perspective of the economic and supply-side risks that affect drug utilization.

C. Practical Implications for Healthcare Stakeholders

The outcomes of this research have real-world ramifications for a diverse group of individuals and organizations that are involved in or impacted by the healthcare system. Hospital procurement departments can utilize risk scores of usage to guide their decisions on the inclusion in the formulary by pointing out those compositions which may create issues with affordability or supply. In addition, the strategies for procurement can be refined by the selection of compositions that have a stable competition and availability.

For the doctors, the prescription support system provides them a data-oriented way to recognize the cost-effective and non-exhausted alternatives that are of the same chemical composition. This kind of support can help in cutting down the unnecessary costs for patients while keeping the drugs equivalent therapeutically, mostly in the outpatient and primary care settings.

From the perspective of policy, the examination of the utilization risk can be the basis for interventions concerning regulatory and pricing. Government can employ risk indicators to spot compositions attracting price monitoring, suppliers' diverse sourcing, or being included under price control schemes. The outlined procedure can also be employed in public health procurement programs and deciding insurance reimbursement planning.

D. Limitations of the Study

Although the proposed framework has its advantages, it also has several limitations which need to be considered. The analysis, for instance, is performed on market-based medical data only and does not take into account clinical outcomes, patient compliance, or the doctor's intention. Hence, the utilization risk scores mirror economic and supply-related issues, not the benefits or safety of the drug interventions.

Moreover, the data used is not the complete picture of the pharmaceutical market and represents a situation at a particular moment in time. The pharmaceutical market is in constant flux with pricing, availability, and manufacturer participation changes happening all the time. A longitudinal study would

TABLE IV
EXAMPLE OF PRESCRIPTION SUPPORT VIA SUBSTITUTE IDENTIFICATION

Brand	**Price**	**Availability**	**Recommendation**
Brand X	High	Available	Not Recommended
Brand Y	Moderate	Available	Recommended
Brand Z	Low	Available	Highly Recommended

be necessary to show the effect of time on trends and the risk of drug misuse.

In addition, the model being unsupervised in nature restricts its direct validation against true outcomes. Despite domain-driven explanation lending credibility to the results, validation and interpretation could be improved by expert review or partial labeling being included in the future work.

To conclude, the present framework is limited to the evaluation of single compositions and does not take into consideration multi-drug treatment regimens or therapeutic substitution among different chemical classes, which are, however, common in practice.

E. Ethical and Transparency Considerations

Using only market-based data guarantees adherence to ethical research standards and reduces to the minimum the likelihood of privacy concerns related to patient-level data. Nevertheless, the openness of model interpretation is still very crucial to stop the misuse or over-dependence on automated risk scores.

The transparency of engineered features and anomaly-based scoring plays a major role in responsible deployment by making it possible for the stakeholders to comprehend the reasons behind the risk of utilization. Communication about the model's limitations should be made very clear so that the outputs will be utilized as decision-support tools and not as strict rules.

VIII. CONCLUSION AND FUTURE WORK

Employing solely market-driven data ensures compliance with the most ethical research practices and at the same time cuts down to almost zero the possibility of privacy issues connected with patient data. However, the clarity of interpretation of the model is still very important to prevent the misuse or over-reliance on automated risk scores.

The transparency of feature engineering and anomaly detection scoring is a major factor in responsible deployment as it allows the stakeholders to understand why the utilization of the risk was considered. It is recommended to have very clear communication about the limitations of the model so that the outputs are used as decision-support tools rather than being treated as strict rules.

Future research is going to be concentrating on the modification of the developed framework by adding longitudinal market data which would help in recognizing the temporal trends in the factors like pricing, availability, and manufacturer participation. The collaboration with prescription-level or clinical outcome data, if available, would not only allow better validation but also make the analysis of the utilization richer. More studies might also investigate the possibility of applying the framework to multi-drug treatment regimes and therapeutic substitution among different chemical classes, thus offering a complete picture of the dynamics of drug utilization.

REFERENCES

[1] S. Shalini, "Drug utilization studies – An overview," *Indian Journal of Pharmaceutical Sciences*, vol. 72, no. 2, pp. 149–156, 2010.

[2] A. Gujar, V. Gulecha, and A. Zalte, "Drug utilization studies using WHO prescribing indicators from India: A systematic review," *Health Policy and Technology*, vol. 10, no. 3, p. 100547, 2021.

[3] N. Goruntla *et al.*, "Evaluation of rational drug use based on WHO/INRUD core drug use indicators," *Journal of Clinical and Diagnostic Research*, vol. 17, no. 3, pp. FC01–FC06, 2023.

[4] R. Joshi *et al.*, "Assessment of prescribing pattern and completeness as per WHO core drug use indicators," *Indian Journal of Pharmacology*, vol. 54, no. 5, pp. 361–367, 2022.

[5] A. Ray, "A cost variation analysis of drugs available in the Indian market," *Postgraduate Medical Journal*, vol. 96, no. 1138, pp. 491–496, 2020.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422.

[7] L. K. Vora *et al.*, "Artificial intelligence in pharmaceutical technology and drug delivery," *Pharmaceutics*, vol. 15, no. 7, p. 1916, 2023.