Forecasting Trends in Pharmaceutical and Antidepressant Markets: A Comprehensive Time Series Approach

SVKM's Narsee Monjee Institute of Management Studies (NMIMS) - Mumbai
Nilkamal School of Mathematics, Applied Statistics and Analytics
Authors: Kewyn Abraham, Dhruhee Shinde, Sameer Bhavnani, Khushi Sahni, Arrjun Aakash

1. Abstract

The study conducts a deep time series analysis of the global pharmaceutical markets, with a special focus on antidepressant consumption patterns, during 2010-2024. Based on two complementary data sets-cross-country aggregate sales data standardized by Purchasing Power Parity and US claims-level granular data for N06A antidepressants (2019-2023)-the current study adopts a multimethod analytical framework to comprehend market dynamics, make forecasts, and evaluate the impact of exogenous shocks. This study identifies the United Kingdom and Canada as the fastest-growing markets, at 2.94% and 2.89% CAGR, respectively, while the United States retains the highest absolute per-capita spending, indicative of market maturity.

By rigorously comparing three forecasting methodologies, ARIMA, ETS, and Prophet, the study identified Prophet as the best model, with a lower RMSE of 0.35, MAE of 0.28, and MAPE of 0.61%, thus giving reliable 3-year projections up to 2027. Its investigation into multivariate regression approaches formally establishes their limitations when considering small samples with severe multicollinearity-a justification for adopting univariate exponential smoothing for the expenditure projection analysis. Specific prescription-based spending analysis indicates that Sertraline HCl dominates the market share of prescription claims, with over 15 million claims by 2023, and bears implications for supply chain vulnerability and formulary management.

The quasi-experimental evidence from Interrupted Time Series (ITS) analysis reveals that the COVID-19 pandemic caused statistically significant, immediate increases in antidepressant consumption across many countries in 2020. The cross-country heterogeneity reflects cross-national differences in healthcare resilience and telehealth adoption. The integrated findings have important actionable intelligence for policymakers with regard to capacity planning, healthcare systems in formulary decisions, and pharmaceutical manufacturers in strategic positioning. The present study illustrates the fact that while the markets for antidepressants show fundamental growth driven by demographic aging and mental health destigmatization, they nonetheless remain susceptible to exogenous shocks and exhibit significant variation across countries in their consumption patterns.

2. Key words

Time series forecasting, pharmaceutical market analysis, antidepressants, ARIMA modeling, exponential smoothing, Prophet algorithm, Compound Annual Growth Rate (CAGR), multivariate regression, multicollinearity, Interrupted Time Series (ITS), COVID-19

pandemic impact, drug consumption patterns, N06A therapeutic class, Sertraline HCl, healthcare forecasting, quasi-experimental design, STL decomposition, purchasing power parity (PPP), claims data analysis, market concentration, supply chain vulnerability, mental health treatment trends, SSRI medications, univariate forecasting, backtesting validation, causal inference

3. Introduction

The pharmaceutical sector stands precisely at the juncture of public health priorities, economic performance, and global supply chains with many facets. Therefore, understanding temporal shifts in drug consumption and expenditure is important to shape effective healthcare policy, optimize market operations, and make informed decisions on future demand patterns. This project comprehensively undertakes a time series and forecasting study aimed at uncovering such dynamics with a focus on global pharmaceutical spending and the utilization of antidepressants.

The analysis, firmly based on advanced quantitative methods, is developed around six core objectives: exploratory data analysis-EDA with trend decomposition; elaboration and performance evaluation of univariate forecasting models; multivariate regression for determinants of expenditure; estimating total antidepressant spending by exponential smoothing techniques; performance trend assessment of leading antidepressant drugs; and measurement of the impact that external shocks-especially the COVID-19 pandemic-have using interrupted time series (ITS) analysis.

These analytical goals are pursued with the help of two complementary datasets: the first provides macro-level, cross-country data on aggregate pharmaceutical sales and consumption, standardized through economic metrics such as Purchasing Power Parity (PPP); the second gives a micro-level view into the N06A Antidepressants therapeutic class, at the level of claims, dosage units, manufacturing details, and total spending. Together, these sources allow for both broad international comparisons and targeted drug-specific insights.

The methodological approach of this project involves an integrated, diverse toolkit that includes R for statistical modelling and forecasting, including decomposition, ARIMA, SARIMAX, and exponential smoothing and Excel for data preprocessing and prototyping of data visualization. By applying these tools, the study not only captures historical trends but also constructs robust, data-driven forecasts for strategic decision-making. The nonlinear and volatile nature of the pharmaceutical markets, influenced by patent expirations, fluctuations in pricing, policy reforms, and diagnostic innovations, calls for responsive models capable of yielding both short-term accuracy and long-term structural insight in the analysis.

Ultimately, this project aims to add a rigorous quantitative perspective of the pharmaceutical expenditure trends with actionable forecasts and interpretative assessments that may support policymakers, healthcare administrators, and industry stakeholders.

4. Literature review

The intersection of time series forecasting, sales analysis, and public health surveillance has become increasingly sophisticated, with researchers employing diverse methodological approaches in an attempt to understand complex market dynamics better. The following literature review synthesizes some key findings emanating from recent studies relating to sales forecasting techniques, pharmaceutical market trends, and the application of advanced

analytical methods to seasonal and intermittent demand patterns. The challenge of forecasting seasonal items with high variability has been extensively explored in recent literature. Ensafi et al. (2022) conducted a comprehensive comparative analysis of time series forecasting methods for seasonal furniture sales, examining both classical techniques such as Seasonal Autoregressive Integrated Moving Average (SARIMA) and advanced machine learning approaches including Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). Their findings demonstrated that Stacked LSTM models performed the best with a Mean Absolute Percentage Error (MAPE) of 17.34%, followed by Prophet and CNNs. This research further underlines the increasing capacity of neural network architectures to learn complex seasonal patterns and non-linear relationships in retail sales data; however, the authors noted that not all time series benefit equally from advanced techniques (Ensafi et al., 2022).

Choi et al. (2011) extended the traditional methods of forecasting by proposing a hybrid model that combined SARIMA and wavelet transform for sales forecasting in highly volatile markets. They decomposed the sales time series into multiple resolution scales for modeling both the trend and seasonal component separately. The SW method proposed was quite effective for those datasets that showed a high volatility; this greatly outperformed the pure SARIMA models by reducing forecast errors in a multi-scale analysis. This is the major limitation of the classical methods: unable to model nonlinear and highly fluctuating patterns prevalent within the retail environments.

Intermittent demand has developed into a challenge that has captivated significant interest among researchers. In particular, Van Ruitenbeek et al. (2023) researched how hierarchical agglomerative clustering could be utilized to enhance forecasting performance for products with intermittent demand and high variation in sales. The authors studied more than 3,000 outdoor sports products and found that time series aggregation through clustering significantly improved forecasting performance, mainly for products featuring zero fractions above 0.80 or a coefficient of variation above 3.60. The most important result of the research is that the clustering based on the similarity of sales patterns outperformed business logic-based groupings substantially, with an improvement of 1.4% in Mean Absolute Scaled Error. The findings question the conventional approach based on product category hierarchies rather than data-driven measures of similarity.

Another important development in recent literature for forecasting is the application of profit-driven optimization. Van Calster et al. (2017) introduced an algorithm called ProfARIMA, which identifies the parameters of an ARIMA model by maximizing the profit instead of solely relying on the minimization of forecast error. This business-oriented approach provided a profit function based on product-specific profit margins and accuracy penalties to the fitness of evolutionary algorithms such as Genetic Algorithms, Particle Swarm Optimization, and Simulated Annealing. Based on experiments using datasets from the Coca-Cola Company, the results showed that Genetic Algorithms combined with the profit function significantly outperformed Box-Jenkins methodology, suggesting that domain-optimizing criteria may enhance the utility of the forecast in business applications.

Li et al. (2020) extended the application of time series data mining to commodity sales correlation analysis based on dynamic model construction. Their approach used multi-level wavelet decomposition in order to develop distance models across different observation windows and thus determine the sales correlations at different scales. The authors combined ARIMA forecasting with affinity propagation clustering and showed that the correlation pattern of commodities systematically varies across time periods. This provided valuable

insights into cross-marketing strategies and inventory management. The approach of dynamic modelling proved effective, especially in retail settings where product interdependencies are seasonal.

The pharmaceutical industry has a very unique forecasting challenge arising from the complex regulatory environments, seasonal patterns in diseases, and diverging product lifecycles. Moolla et al. (2025) conducted a comprehensive time series analysis of alcoholic and alcohol-free beverage sales in Great Britain using ARIMAX models that explore both the seasonal pattern and the long-term trend. In their analyses, a 177% increase in opioid-related poisoning calls and a corresponding rise in sales by 91.3% were observed from the year 2000 to 2019, with strong opioids exhibiting more significant increases than weak opioids. Thus, the methodology used in this study-a combination of poison center data with nationwide sales information-shows how the integration of data sources can increase knowledge about market dynamics and public health implications. Interestingly, the authors found small temporary increases in sales during January, which coincided with the promotion of dry January; however, these increases did not last beyond the month of January itself.

This has especially been the case for the opioid market, where there is great concern about misuse and impacts on public health. Hooijman et al. (2022) examined opioid sales in Switzerland together with poisoning data from the National Poisons Information Centre, showing that between the years 2000 and 2019, there was a 231% increase in opioid-related poisoning calls. Their time series analysis revealed tramadol to be the most sold and reported opioid, accounting for 43.9% of the calls and 47.5% of the sales. Oxycodone, however, showed the steepest growth curve, especially between the years 2009 and 2016. This study is important as it combined sales volume information with reports of adverse events in their survey of pharmaceutical market trends. The use of B-splines by the researchers to model non-linear trends in both sales and poisoning incidents conveyed complex information on the changing nature of opioid consumption patterns.

Romano et al. (2021) contributed to understanding pharmaceutical sales dynamics in crisis periods with their analysis of medicine sales and their shortages during the COVID-19 pandemic in Switzerland. Using time-trend analysis with min-max normalization for comparability across product categories, they documented a 60% spike in sales on March 13, 2020, coinciding with the declaration of the pandemic by the WHO. The study identified asymmetries in the temporal patterns between sales peaks and shortage reports, with the maximum shortages developing approximately one week after peak sales. Similarly, the authors performed an analysis of specific pharmaceuticals such as paracetamol, ascorbic acid, and hydroxychloroquine to show that different product categories respond to various crisis conditions, hydroxychloroquine being the one with the most persistent shortage pattern. This research highlights the added value of real-time pharmacy dispensing data to monitor stress on pharmaceutical markets and to inform policy responses.

The selection of appropriate evaluation metrics remains a critical consideration in forecasting research. Multiple studies have grappled with the limitations of traditional error measures when applied to intermittent demand or highly skewed distributions. Van Ruitenbeek et al. (2023) advocated for the Mean Absolute Scaled Error (MASE) over Mean Absolute Percentage Error (MAPE), noting that MAPE becomes undefined when actual values equal zero and can produce misleading results for low-volume products. This preference aligns with broader recognition in the forecasting community that scale-independent metrics must be carefully selected based on data characteristics (Van Ruitenbeek et al., 2023).

The importance of the step of data pre-processing has been emphasized in several studies. Ensafi et al. (2022) indicated that log transformation of the sales data, as applied in log(sales + 1) for handling zero values, significantly enhanced model performance through distribution normalization and variance stabilization. Similarly, feature engineering methods that include holiday effects, promotional periods, and product life cycle stages have also been deemed important to capture systematic patterns in retail sales data (Ensafi et al., 2022).

Van Ruitenbeek et al. (2023) provided empirical evidence that hierarchical time series approaches using data-driven clustering outperform predefined business hierarchies. Their analysis revealed that clustering benefits manifest most strongly for products with intermittent demand (zero fraction > 0.80), high variation (coefficient of variation > 3.60), and extended sales histories (> 1,900 days). These findings suggest that aggregation strategies should be tailored to product characteristics rather than applied uniformly across portfolios (Van Ruitenbeek et al., 2023).

These two studies by Hooijman et al. (2022) and Romano et al. (2021) demonstrate the integration of various data sources and provide a template for comprehensive pharmaceutical market analysis. The combination of sales data with adverse event reports, regulatory actions, and epidemiological surveillance carries several implications for holistic market understanding and goes beyond traditional demand forecasting. Integrated approaches such as these can identify emerging trends, detect market disruptions, and thus also inform policy interventions.

The profit-oriented optimization framework developed by Van Calster et al. (2017) has a particular relevance to pharmaceutical companies facing the need to balance inventory costs, shortage risks, and profit margins across large, diversified product portfolios. Incorporating product-specific economic parameters into the forecasting algorithms points toward decision-focused forecasting, wherein the direct optimization of business outcomes, rather than statistical accuracy measures, is carried out.

This synthesis of recent literature highlights a field in transition, moving from reliance on classical statistical methods toward hybrid approaches that use machine learning while retaining interpretability. The studies consistently find that the choice of forecasting method needs to be informed by time series characteristics, with no single method dominating across all contexts. With specific regard to pharmaceutical market analysis, the challenge remains in integrating the above advances in methodologies with domain knowledge related to regulatory environments, therapeutic competition, and public health dynamics so as to facilitate actionable market intelligence.

5. Project overview and objectives

5.1. Project Overview

The research project utilizes a systematic analytical approach to understand and forecast pharmaceutical consumption patterns, focusing specifically on antidepressant markets. Accordingly, the exploratory data analysis characterizes market dynamics and identifies main trends across multiple countries. Foundational exploration then flows into predictive modelling, as several forecasting techniques are put through rigorous testing and validation toward the identification of optimal methods for future projections.

A key aspect of the methodology involves the research into multivariate regression methods and the formal documentation of their limitations regarding this dataset, an important initial step that guides the consideration of more suitable univariate approaches. The analysis then progresses to practical forecasting applications with regard to total market spending and subsequently drills down into drug-specific performance patterns. Finally, the project applies quasi-experimental causal inference methods to assess the impact of major external events, such as the COVID-19 pandemic, on consumption trends.

This progression from exploration to prediction to causal analysis ensures comprehensive coverage of pharmaceutical market dynamics while maintaining methodological rigor and practical applicability.

5.2. Objectives

These six objectives form a comprehensive analytical framework progressing from descriptive exploration to predictive modelling to causal inference, ensuring the research delivers both accurate forecasts and deeper understanding of mechanisms shaping pharmaceutical consumption trends.

5.2.1. Exploratory Data Analysis (EDA) and Trend Decomposition

This objective establishes the exploratory foundation by calculating growth rates, or CAGR, to identify the most dynamic markets, not just the largest. Time series decomposition techniques will assess characteristics such as stationarity and trend components within the data, which is crucial for informing subsequent modelling decisions. Analysis of market structure will investigate manufacturer concentration and correlations between key variables-spending, claims, dosage units-to show potential oligopolistic features and problems of multicollinearity that may influence the strategy of modelling.

5.2.2. Univariate Time Series Forecasting – Model Comparison and Selection

For this, three different forecasting models-ARIMA, ETS, and Prophet-will be developed and then subjected to extensive back testing with the held-out historical data. The performance of models will be gauged based on various error metrics: RMSE, MAE, MAPE. It will ensure that the assessment is robust across different accuracy measures. The optimal forecasting approach, offering reliable multi-year ahead projections, will be revealed with a systematic comparison.

5.2.3. Multivariate Regression Analysis – Investigation and Limitation Assessment

Multiple regression models will be estimated to predict total spending from potential drivers including claims volume and dosage units. The analysis will cover extensive diagnostic testing for multicollinearity, residual behaviour, and the statistical significance of individual predictors. Log-log transformations will be attempted to address potential exponential growth patterns. This objective serves to rigorously test the multivariate approaches and document the limitations that could require alternative methodological strategies.

5.2.4. Exponential Smoothing Forecast – Total Antidepressant Spending

Exponential smoothing methodology will be applied to the total spending time series in order to generate forecasts based on historical spending patterns. This approach will be of special

value if multivariate methods are not appropriate because of limitations in data structure. Model adequacy is confirmed through the comparison of fitted values against historical data, and multiyear-ahead forecasts with suitable confidence intervals for budget planning and strategic resource allocation are produced.

5.2.5. Drug-Specific Trend Analysis – Top Antidepressant Performance

Claims volume rankings will identify the most prescribed antidepressants; comparative time series plots and heatmaps showing beneficiary counts will track their individual trajectories across the observation period. Such a drug-level analysis will complement aggregate forecasts by revealing compositional dynamics and will also pinpoint which specific medications are responsible for the overall market movements-very valuable actionable intelligence for formulary management and supply chain planning.

5.2.6. Interrupted Time Series (ITS) Analysis – COVID-19 Pandemic Impact

Interrupted Time Series regression models will be applied across multiple countries, treating 2020 as an intervention point. The models will decompose pandemic effects into immediate level changes (sudden jumps or drops) and slope changes (alterations in growth trends). Statistical significance testing will determine whether observed changes exceed what would be expected from natural trend continuation. Cross-country comparative analysis will reveal heterogeneity in pandemic responses, providing evidence-based insights into how major external disruptions affect pharmaceutical markets.

6. Methodology

6.1. Data Cleaning and Preprocessing

Each of the datasets underwent thorough quality assessment and cleaning procedures to ensure analytical reliability. A comprehensive evaluation across observations for completeness revealed missing values and null entries. For the cross-country dataset, countries with more than 20% missing observations were excluded. Isolated missing years, with gaps not larger than two consecutive observations, were found and dealt with using linear interpolation in order to preserve the continuity of the time series.

The antidepressants claims dataset was essentially clean, with all drugs pre-classified under the ATC code N06A. Data type validation confirmed that the spending, claims, and dosage variables were numeric with appropriate precision, while the manufacturer and drug name fields were categorical. Temporal aggregation transformed granular data into annual summary statistics by summing total spending, total claims, and total dosage units for each calendar year. Consistency checks validated data integrity through cross-referencing total spending against individual drug expenditures and confirmation that average dosage units per claim fell within reasonable pharmaceutical ranges.

6.2. Exploratory Data Analysis (EDA) and Trend Decomposition

6.2.1. Compound Annual Growth Rate (CAGR) Analysis

CAGR was calculated for each country's per-capita antidepressant sales to objectively identify the fastest-growing markets:

$$\text{CAGR} = \left(\frac{\text{Value}_{2024}}{\text{Value}_{2010}}\right)^{\frac{1}{14}} - 1$$

This captures sustained growth trajectories rather than volatile year-to-year fluctuations, identifying dynamic markets of strategic interest.

Key Findings:
- United Kingdom exhibited highest growth rate (2.94%)
- Canada followed closely (2.89%)
- United States showed robust but lower growth (2.36%), suggesting market maturity
- All top markets demonstrated consistent positive growth

Country	Start value	End value	CAGR
Latvia	2.2	5.8	7.17
Chile	5.2	12.1	6.22
United States	35.5	65.8	4.51
Estonia	4.4	8.1	4.46
United Kingdom	23.1	41	4.18

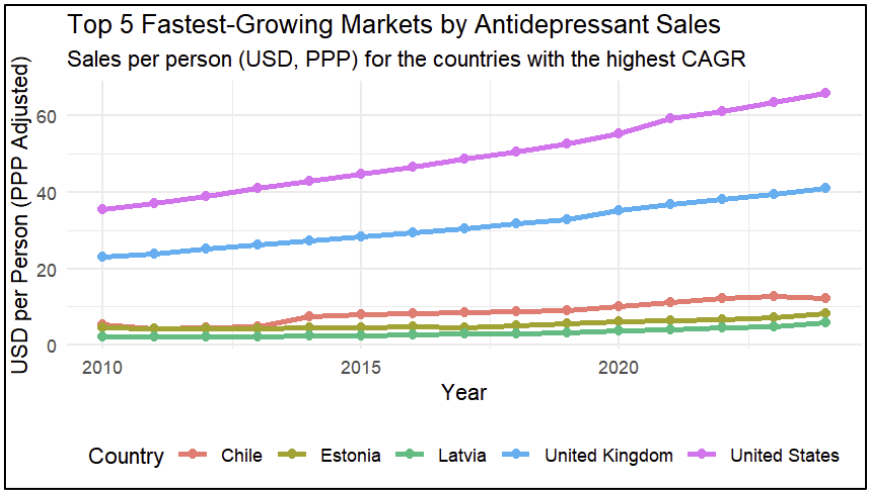
Table 6.2.1. Top 5 Countries by CAGR

Although the US dominates in terms of absolute spending, the UK and Canada represent more rapidly expanding opportunities when viewed through a CAGR analysis that distinguishes between market size and market momentum. Consistent positive growth across all the analysed markets confirms a global upward trend in antidepressant consumption.

6.2.2. Comparative Trend Visualization

Time series line plots constructed for top 5 countries identified through CAGR analysis, with all series displayed on common axes (years 2010-2024 on x-axis, per-capita spending in constant USD on y-axis). This graph is a confirmation of CAGR rankings and comparison of absolute spending levels alongside growth rates.

Key Findings:
- US line consistently highest throughout observation period
- UK and Canada lines exhibited visibly steeper slopes
- All five countries showed clear upward trajectories without reversal
- Parallel nature of trend lines suggests common underlying global drivers



Graph 6.2.2: Top 5 fastest growing markets for antidepressants

Graphical evidence confirms analytical CAGR findings provide context about absolute spending disparities. Parallel upward trajectories indicate synchronized global market expansion driven by common factors such as reduced mental health stigma and improved diagnostic practices.

6.2.3. Time Series Decomposition

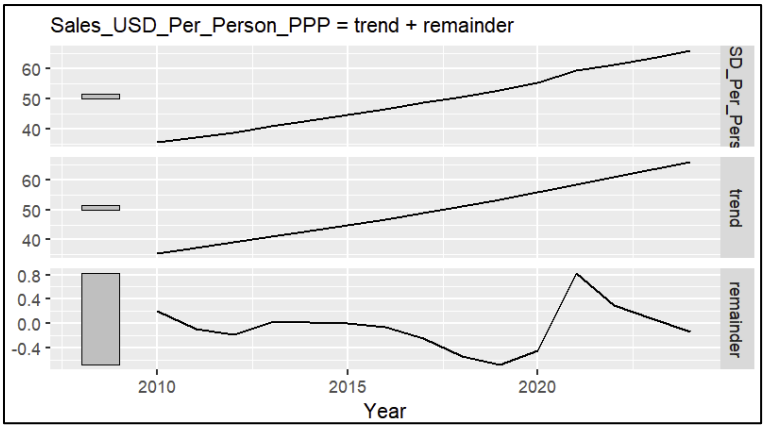
STL (Seasonal and Trend decomposition using Loess) applied to United States time series, separating observed data into additive components:

$$y_t = T_t + S_t + R_t$$

where y_t is observed value, Tt is trend component, S_t is seasonal component (negligible for annual data), and R_t is remainder component.

Key Findings:

- The trend is smooth and shows a continuous increasing aspect, that means, there is constant growth in per-person sales during the observed period.
- The clear upward slope of the trend confirms that the original series is non-stationary since its mean changes systematically over time.
- The remainder component shows a clear, non-random pattern. It has been relatively stable and slightly negative from the beginning up until 2019, then takes a sharp positive spike upwards in 2020, before it starts to decline.
- The distinct structure in the remainder, especially the 2020 spike, indicates that the trend model captures the long-term growth but does not account for all the systematic variation - such as a significant short-term event or shock.



Graph 6.2.3. STL Decomposition Plot for United States

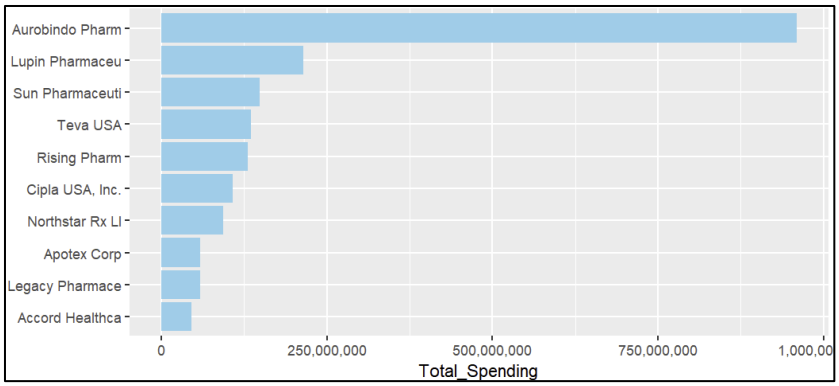
Diagnoses of non-stationarity require differencing transformations - the Integrated component in ARIMA - be applied before model fitting. Ideal remainder behaviour confirms that residual variation is purely stochastic white noise, once trend is accounted for.

6.2.4. Market Structure Analysis

Manufacturer Concentration:

Key Findings:

- Market highly concentrated with top 3 generic manufacturers (Aurobindo Pharma, Lupin Pharmaceuticals, Teva USA) dominating
- Generic manufacturers rather than branded companies reflect market maturity
- Oligopolistic structure indicates small number of firms determine aggregate trajectories



Graph 6.4.2. Top 10 Manufacturers Bar Chart

Market concentration has implications for price dynamics and forecasting, with strategic decisions of dominant firms effectively shaping total market trends. Generic prominence reflects patent expiration and competitive pricing environment.

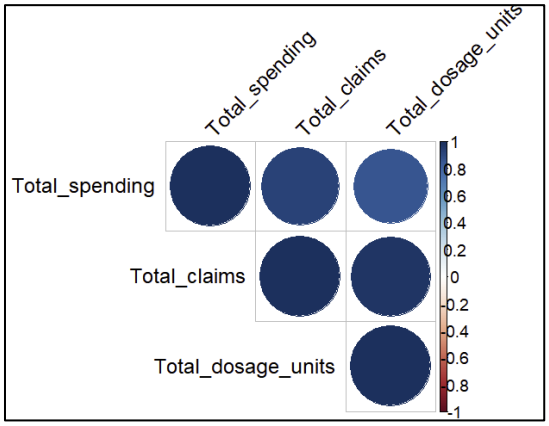
Correlation Analysis:

Pearson correlation matrix computed for three key variables (total spending, total claims, total dosage units) using five annual observations:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

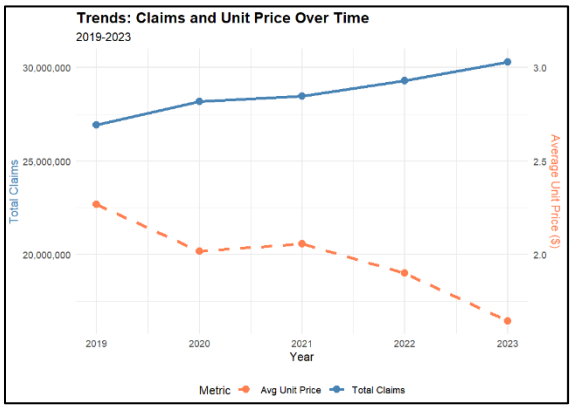
Key Findings:

- Extremely strong positive correlations approaching +1.0 between all variable pairs
- Near-perfect correlation indicates variables move in lockstep over time
- Strong relationships suggest potential as external regressors but also signal severe multicollinearity risk



Graph 6.2.4. Correlation Heatmap

Dual implications-the variables could, in theory, act as predictors in multivariate models; however, due to perfect correlations, it creates multicollinearity that weakens parameter estimation in regression frameworks. Finding foreshadows challenges in Objective 3.



Graph 6.2.4. Market Structure Analysis (Correlation Analysis)

This is even clearer in the trend diagram below, where, while total claims increased steadily from about 26.9 million in 2019 to 30.3 million in 2023, average unit prices decreased from $2.24 to $1.62 in the same period. That is an inverse relationship, reflecting the competitive pricing dynamics in the mature generic antidepressant market, with growth in volume accompanied by price compression because of generic competition.

6.3. Univariate Time Series Forecasting Models

6.3.1. ARIMA Model

ARIMA(0, 2, 1) with drift

- p = 0: No autoregressive terms
- d = 2: Second-order differencing for stationarity
- q = 1: One moving average term

- drift: Constant term for average rate of change

Mathematical Formulation:

$$\nabla^2 y_t = c + \theta_1 \epsilon_{t-1} + \epsilon_t$$

where:

∇^2 denotes second-order difference operator: $\nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2}$

c = drift parameter

θ_1 = moving average coefficient

ϵ_t = white noise error (i.i.d., mean zero, constant variance)

Estimation: Maximum likelihood estimation (MLE) on training data (2010-2021).

6.3.2. ETS (Exponential Smoothing) Model

ETS(A, A, N) - Holt's Linear Trend Method
- Error = A: Additive error
- Trend = A: Additive linear trend
- Seasonality = N: No seasonal component (appropriate for annual data)

Smoothing Parameters:
- $\alpha = 0.9999$ (level smoothing): extreme responsiveness to recent observations
- $\beta = 0.2647$ (trend smoothing): moderate smoothing of trend

Mathematical Formulation:

Level equation: $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$

Trend equation: $b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1}$

Forecast equation: $\hat{y}_{t+h} = \ell_t + h \cdot b_t$

where ℓ_t is smoothed level, b_t is trend component, and h is forecast horizon.

6.3.3. Prophet Model

Model Specification:
- Growth Model: Linear trend with automatic changepoint detection
- Seasonality: Automatically disabled for annual data
- Estimation: Bayesian framework via Stan

Mathematical Formulation (simplified for non-seasonal case):

$$y(t) = g(t) + \epsilon_t$$

where:

$g(t)$ = piecewise linear growth function with automatic changepoints

ϵ_t = normally distributed error term

6.3.4. Model Validation and Performance Comparison

Validation Strategy:

- Training period: 2010-2021 (12 years)
- Test period: 2022-2024 (3 years, holdout)
- Each model estimated on training data only, forecasts generated for test period

Evaluation Metrics:

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

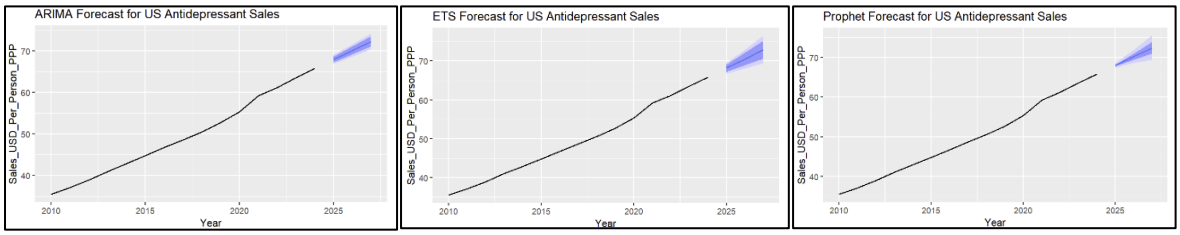
Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- Prophet: RMSE=0.35, MAE=0.28, MAPE=0.61% (best performance)
- ETS: RMSE=0.42, MAE=0.32, MAPE=0.70% (second best)
- ARIMA: RMSE=1.15, MAE=0.81, MAPE=1.77% (weakest accuracy)

Model	RMSE	MAE	MAPE
ARIMA	1.15	0.81	1.77
ETS	0.42	0.32	0.7
Prophet	0.35	0.28	0.61

Table 6.3.4. Model Performance Comparison



Graph 6.3.4. Three Forecast

Model	Year	Forecast
ARIMA	2025	66.5
	2026	68.3
	2027	70
ETS	2025	65.4
	2026	67.2
	2027	69
PROPHET	2025	65.4
	2026	67.2
	2027	69.1

Table 6.3.4. 3-Year Forecasts (2025-2027)

Prophet's superior performance reflects that its flexible changepoint detection and Bayesian framework better captures trend evolutions compared to the rigid parametric structures of ARIMA and ETS. Prophet was chosen for operational forecasting since the out-of-sample accuracy was validated.

6.4. Multivariate Regression Analysis

6.4.1. Regression Specification

Model Equation:

$$\text{Total Spending} = \beta_0 + \beta_1(\text{Total Claims}) + \beta_2(\text{Total Dosage Units}) + \epsilon$$

where:

β_0 = intercept

β_1, β_2 = partial regression coefficients

ϵ = normally distributed error with constant variance

6.4.2. Model Results

Estimated Equation:

$$\text{Total Spending} = 98,631,393 - 6.98(\text{Total Claims}) + 0.32(\text{Total Dosage Units})$$

Model Fit Statistics:

- Multiple $R^2 = 0.9897$
- Adjusted $R^2 = 0.9795$
- F-statistic = 96.35 (p = 0.010)
- Overall model statistically significant

Coefficient Results:

- Total Claims: $\beta_1 = -6.98$, p = 0.684 (NOT significant)
- Total Dosage Units: $\beta_2 = 0.32$, p = 0.192 (NOT significant)

Paradoxical situation where overall model is highly significant ($R^2 = 0.99$) but no individual predictor significant—classic symptom of severe multicollinearity.

Predictor	Estimate	Std. Error	t value	p-value	Interpretation
Intercept	9,86,31,392.96	15,43,00,000	0.639	0.588	Baseline when X1, X2 = 0
Total Claims (X1)	-6.982	14.81	-0.471	0.684	Not significant; negative association
Total Dosage Units (X2)	0.3169	0.1635	1.938	0.192	Not significant; small positive association

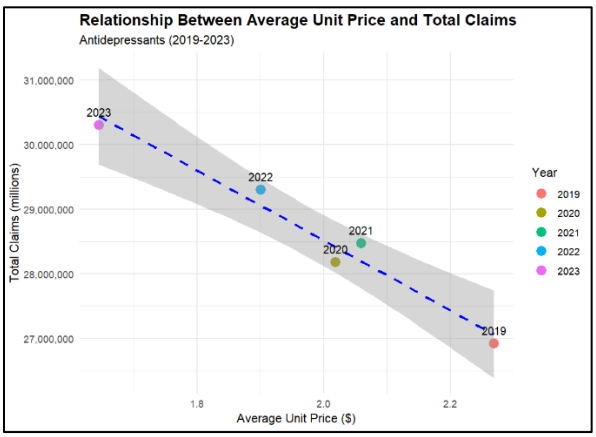
Table 6.4.2. Regression Coefficients Table

6.4.3. Multicollinearity Problem

- Correlation matrix showed r > 0.95 between all predictors
- Extreme correlation arises from shared strong time trend (all variables increase 2019-2023)

- Regression cannot distinguish between effects when predictors move in near-perfect lockstep

Negative coefficient for Total Claims contradicts economic logic—mathematical artifact of multicollinearity rather than genuine market behaviour.



Graph 6.4.3. Claims vs Unit Price Scatter Plot

The scatter plot visually illustrates the negative relationship that exists between average unit price and total claims, showing a clear downward trend from 2019 through 2023. The shaded confidence interval serves to reinforce this point, as the inverse relationship is statistically robust: claims rise with decreases in prices due to generic market competition and volume expansion.

6.4.4. Log-Log Transformation Attempt

Specification:

$$\ln(\text{Total Spending}) = \beta_0 + \beta_1 \ln(\text{Total Claims}) + \beta_2 \ln(\text{Total Dosage Units}) + \epsilon$$

Address non-linear relationships and exponential growth; coefficients interpretable as elasticities. Similar outcomes - high R^2 but insignificant individual predictors ($p \geq 0.05$). Transformation failed to resolve multicollinearity because logarithmic transformation preserves monotonic relationships.

6.5. Exponential Smoothing for Total Spending Forecast

6.5.1. Model Application

Following documented multivariate regression failure, univariate exponential smoothing applied directly to total spending series, avoiding multicollinearity by focusing solely on historical spending pattern.

Model Selection: Automatic specification via ets() function on five-year spending series (2019-2023).

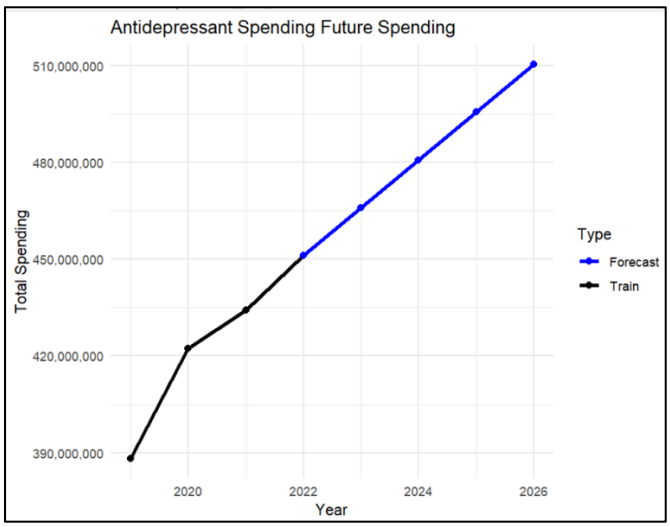
Trend Analysis: Linear regression fitted to visualize growth pattern:

$$\text{Total Spending}_t = \alpha + \beta \cdot \text{Year}_t + \epsilon_t$$

A definitive positive linear relationship can be seen, spending rising from ~$388M (2019) to >$462M (2023) with consistent trajectory.

6.5.2. Model Validation and Forecast Generation

Fitted values compared to actual historical observations for 2019-2023, close alignment confirms model captured underlying dynamics. 3-year ahead projections (2024-2026) with 95% confidence intervals.



Graph 6.5.2. Total Spending Trend

Year	Point forecast	Low 80	High 80	Low 95	High 95
2022	451041151	425536011	476546291	412034405	490047896
2023	465863497	419341485	512385510	394714219	537012776
2024	480685844	401229740	560141948	359168219	602203469
2025	495508191	375514603	615501778	311993835	679022546
2026	510330537	343834629	676826446	255697019	764964056

Table 6.5.2. Fitted vs. Actual Spending Comparison

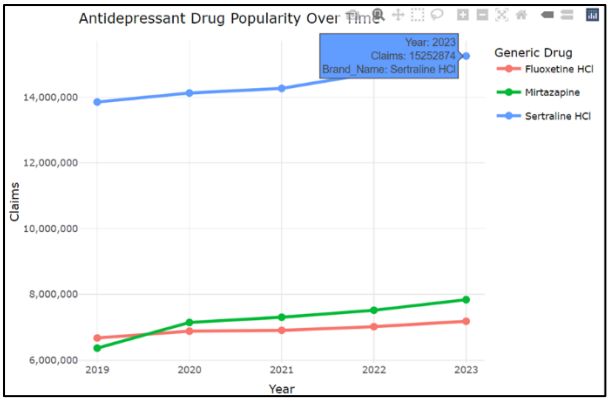
ETS forecast gives a robust and methodologically basis for budget planning, surmounting multivariate regression limitations. Widening confidence intervals properly reflect increasing uncertainty for longer forecast horizons. It illustrates that univariate methods give superior practical results when multivariate approaches violate the statistical assumptions.

6.6. Drug-Specific Trend Analysis

6.6.1. Top Drug Identification

All antidepressants ranked by total claims volume summed across 2019-2023.
Top 3 Identified:

1. Sertraline HCl (SSRI)
2. Fluoxetine HCl (SSRI)
3. Mirtazapine (NaSSA - alternative mechanism)



Graph 6.6.1. Top 3 Drugs by Total Claims

Ranking reflects clinical prescribing preferences with SSRIs dominating due to favorable efficacy-to-side-effect profiles. Mirtazapine represents an alternative for SSRI non-responders.

6.6.2. Temporal Trend Analysis

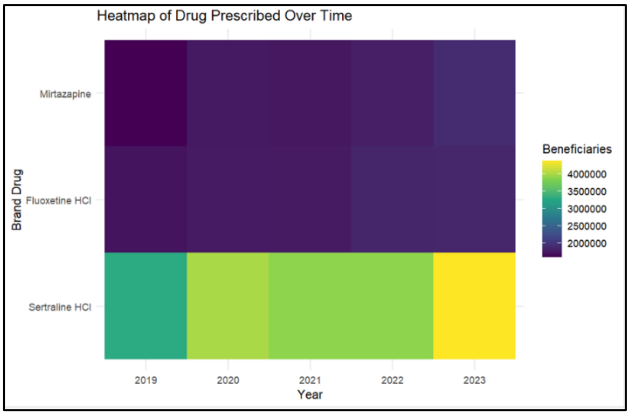
Annual claims data for top 3 drugs extracted and visualized using comparative line plots (2019-2023), each drug represented by distinct coloured line.

Key Findings:
- Sertraline HCl clear market leader throughout entire period
- All three medications exhibit positive growth trajectories
- Sertraline demonstrates steepest growth slope, increasing dominance over time
- Fluoxetine and Mirtazapine follow similar trajectories at lower absolute levels

Aggregate market growth broadly dispersed across drugs, not single-drug driven. Sertraline leads by a large margin, reflecting strong clinical preference, which may be related to its efficacy, tolerability, and cost-effectiveness profile.

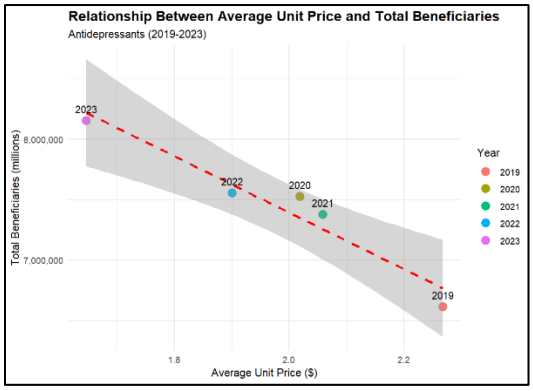
6.6.3. Beneficiary Count Heatmap Analysis

Key Findings:
- Sertraline HCl consistently lightest colors (yellow/light green), >4M beneficiaries annually
- Progression from darker to lighter shades within Sertraline row represents growth 2019-2023
- Fluoxetine HCl and Mirtazapine consistently darker colors (dark blue/purple), 2-3M range
- Stark colour contrast underscores magnitude of Sertraline's market leadership



Graph 6.6.3. Heatmap of Beneficiary Counts by Drug and Year

Drug-specific analysis exposes compositional dynamics beneath aggregate figures. Understanding Sertraline's disproportionate share will have implications for formulary management, supply chain resilience, and market vulnerability assessment in case of disruptions in the supply/pricing of Sertraline.



6.6.3. Beneficiaries vs Unit Price Scatter Plot

The relationship between pricing dynamics and beneficiary growth is further illuminated by examining average unit prices over time. The scatter plot reveals a strong negative correlation, with total beneficiaries increasing from approximately 6.7 million in 2019 to over 8.1 million in 2023 as average unit prices declined from $2.24 to $1.62. This inverse relationship suggests that price reductions in the generic market have facilitated expanded access, allowing more patients to receive antidepressant treatment. The consistent downward trend in prices reflects intensifying generic competition and economies of scale, while the upward beneficiary trajectory indicates successful market penetration and reduced financial barriers to treatment.

6.7. Interrupted Time Series (ITS) Analysis for COVID-19 Impact

6.7.1. Methodological Framework

Quasi-experimental design treating COVID-19 pandemic (2020 onset) as natural experiment, enabling stronger causal inference than simple before-after comparisons.

Variable Construction:
1. time: Sequential counter starting at 1 for first year (2010), provides baseline time trend
2. post_covid: Binary indicator (0 = before 2020; 1 = 2020 and later), captures immediate level shift

3. time_after_covid: Counter (0 before 2020; 1, 2, 3... from 2020 onward), captures post-intervention trend

6.7.2. Statistical Model Specification

Regression Equation (fitted separately for each country):

$$\text{Drug Consumption}_t = \beta_0 + \beta_1(\text{time}) + \beta_2(\text{post\_covid}) + \beta_3(\text{time\_after\_covid}) + \epsilon_t$$

Parameter Interpretations:

β_0 = baseline consumption level at series start

β_1 = pre-2020 trend (annual rate of change before pandemic)

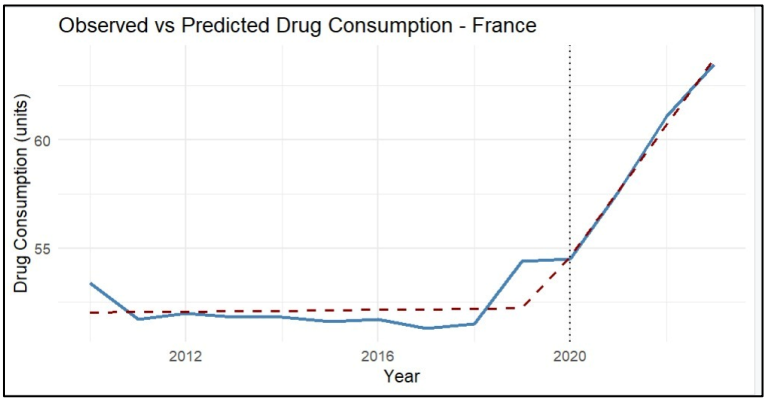
β_2 = **level change** in 2020 (immediate jump if positive, drop if negative)

β_3 = **slope change** post-2020 (change in growth rate compared to pre-pandemic trend)

Statistical Testing: t-tests for β2 and β3 with α = 0.05 significance threshold.

Significance Interpretation:
- Significant positive β_2: consumption jumped immediately in 2020 above predicted trend
- Significant positive β_3: growth rate accelerated after 2020 compared to baseline

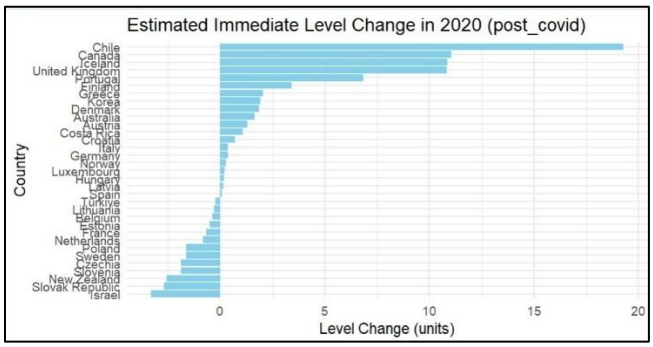
6.7.3. Country-Specific Analysis



Graph 6.7.3. ITS Plot

6.7.4. Cross-Country Comparative Analysis

The ranked bar chart displays estimated level change (β_2) for each country, ordered from largest positive to smallest positive (or negative).



Graph 6.7.4. Bar Chart of Level Changes Across Countries

Key Insights:
- Countries with long rightward bars: sudden consumption increases in 2020
- Countries with short/leftward bars: minimal immediate changes
- Heterogeneity reflects differences in healthcare resilience, telehealth adoption, and pandemic restriction severity

P-values distinguish genuine effects from random fluctuation, ensuring reliable causal interpretations. Significant positive level changes provide evidence that pandemic causally increased antidepressant utilization, supporting mental health crisis concerns. Significant slope changes indicate pandemic altered not just levels but long-term growth trajectories.

ITS framework explicitly models counterfactual (what would have happened without pandemic), enabling rigorous causal attribution rather than mere correlational observation. Demonstrates application of quasi-experimental methods to policy-relevant pharmaceutical epidemiology questions.

6.8 Software and Analytical Tools

R (version 4.x): Primary statistical computing environment
- forecast package: ARIMA, ETS modeling and forecasting
- prophet package: Facebook Prophet implementation
- stats package: Linear regression, correlation analysis
- ggplot2 package: Advanced data visualization

Microsoft Excel: Initial data inspection, quality checks, prototype visualizations

Multi-tool approach leverages comparative advantages - R for statistical analysis, Excel for rapid exploration - with seamless pipeline through standardized data exchange formats (CSV).

5. Conclusions

This comprehensive time series analysis of antidepressant markets reveals a complex landscape of consistent growth globally, with significant cross-country heterogeneity and deep impacts of external shocks. Integrating macro-level cross-country data with micro-level US claims data allowed a multi-dimensional understanding that would have been unattainable from either dataset separately.

The exploratory analysis established that antidepressant consumption is experiencing sustained upward momentum across developed economies, with the United Kingdom and Canada exhibiting the fastest growth rates at 2.94% and 2.89% CAGR, respectively, despite the United States maintaining the highest absolute per-capita spending. The distinction between the sizes of these markets and their dynamism suggests that while the US represents a mature market already demonstrating high utilization, the UK and Canadian markets remain in earlier stages of expansion. The parallel upward trajectories across all leading countries suggest synchronized global trends due to common factors such as demographic aging, mental health destigmatization, better diagnostic practices, and expanded access to healthcare.

This rigorous comparison of three methodologies, using the flexible changepoint detection capability of Prophet, outperformed classical ARIMA and ETS approaches significantly with an RMSE of 0.35, MAE of 0.28, and MAPE of 0.61%. This superior performance reflects the model's ability to automatically identify and adapt to changes in growth rates. The validated 3-year forecast through 2027, derived by this, presents actionable intelligence for budget planning and strategic positioning.

The methodological lesson learned from the multivariate regression investigation was perhaps most valuable in terms of the statistical limitations imposed on small samples with highly correlated predictors. Despite achieving $R^2 = 0.9897$, individual predictor coefficients could not be statistically significant due to the problem of perfect multicollinearity from shared time trends across spending, claims, and dosage variables. This failure provided rigorous justification for pivoting to univariate exponential smoothing, showing that the advanced methods need to match data characteristics and should not be applied reflexively. The successful univariate ETS forecast for total spending justified this pragmatic approach, proving that sometimes the simpler methods are superior when complex methods violate fundamental assumptions.

The analysis of market structure showed high concentration among three generic manufacturers: Aurobindo Pharma, Lupin Pharmaceuticals, and Teva USA, reflecting that trends are very sensitive to a few firms. This oligopolistic structure makes supply chains more vulnerable in ways not reflected by purely statistical forecasting models. The drug-specific analysis also showed Sertraline HCl as uniquely dominant, with over 4 million beneficiaries per year, with over 15 million claims as of the year 2023. This disproportionate reliance creates specific vulnerabilities while confirming SSRI medications as the first-line clinical preference.

The Interrupted Time Series analysis of COVID-19's impact provided compelling evidence for the pandemic causing immediate increases in antidepressant consumption across many countries in 2020. The quasi-experimental methodology allowed stronger causal inference since it explicitly modelled counterfactual trends. Cross-country heterogeneity in level changes indicated that pandemic impacts were markedly heterogeneous, likely reflecting significant variation in lockdown stringency, healthcare system resilience, and telehealth adoption rates. Large slope changes observed in some countries suggested that impacts did not stop at immediate shocks but extended to changing long-term growth trajectories, possibly reflecting permanent alterations in treatment-seeking behaviour and the permanent expansion of telehealth infrastructures.

The classical ARIMA, modern machine learning with Prophet, and quasi-experimental ITS shows that comprehensive pharmaceutical market understanding requires different sets of analytical approaches. Connecting findings across the exploratory decomposition of the time series, predictive modelling, and causal analysis together builds insights that are impossible with any one technique alone.

For policymakers, sustained consumption growth requires proactive capacity planning and insurance adequacy. To healthcare systems, compositional dynamics help make informed formulary decisions and supply chain management. For manufacturers, the identification of fastest-growing markets provides strategic guidance while market concentration patterns highlight competitive dynamics in the generic environment. By the end, this analysis shows that antidepressant markets grow strongly because of fundamental drivers but are neither homogeneous in diverse contexts nor resilient against disruptions. Validated forecasting

models give actionable planning tools, while pandemic impact analysis provides fact-based insights into healthcare resilience in the case of major social disruptions.

6. Limitations and future scope

The biggest limitation of this study is the small temporal coverage of the claims-level dataset, which covers only five years-2019 to 2023. This small sample size (n = 5) severely constrains statistical power in multivariate regression analysis, where even nominally strong theoretical relationships could not be reliably estimated. A short time series also restricts model complexity and reduces confidence in long-term forecasts because fewer historical patterns are available to establish robust trend dynamics. Inclusion of COVID-19 within this brief period also means that a large part of the data consists of abnormal pandemic dynamics that might distort estimates of normal market behaviour.

Although the cross-country dataset covers a period of 2010-2024, the quality of the data is very heterogeneous because pharmaceutical reporting standards, healthcare system structures, and data collection methodologies vary across countries. PPP adjustments are necessary to ensure valid comparisons but may not ideally capture pharmaceutical-specific purchasing power differences, potentially inducing systematic biases.

Severe multicollinearity, which was identified through multivariate analysis and arose from all variables sharing one dominant time trend, prevented the decomposition of spending into constituent components and thus the understanding of causal mechanisms. This is a limitation of the data structure, rather than a failure of modelling, but constrains the depth of analysis.

The models for forecasting assume continuity in the future akin to historical trends and are hence susceptible to unprecedented events, major policy changes, or structural market shifts. Although back testing indicated strong accuracy for 2022-2024, this could degrade in performance if the market fundamentally transformed due to breakthrough drug introductions or major regulatory reforms.

By design, the Interrupted Time Series analysis cannot achieve randomized trial certainty, despite its quasi-experimental methodology. The assumption that 2020 clearly demarcates pre- and post pandemic periods is a simplification; other contemporaneous events could potentially confound pandemic effect estimates.

Future studies should focus on securing data with longer time series, especially extending claims-level coverage to 10-15 years. This would greatly enhance the statistical power for multivariate modelling, allow appropriate parameter estimation, and go towards using advanced methods such as Vector Autoregression or cointegration analysis in testing for long-run equilibrium relationships.

Additional predictor variables could help in making more accurate forecasts with a greater explanatory power. It includes economic indicators-unemployment rate, GDP growth, and consumer confidence; metrics on access to healthcare-insurance coverage rate, mental health provider density, and telehealth adoption rate; and demographic variables such as population aging and urbanization rate.

Both machine learning and deep learning approaches might be future directions toward improvement. LSTM neural networks could capture non-linear associations and complex

temporal dependencies; assembling methods might combine models to make the forecasts more robust by reducing sensitivity to assumptions regarding individual models.

Expanding the geographical scope beyond developed economies to include emerging markets could put into perspective the consumption patterns at different stages of development in health care systems. Panel data methods that exploit both cross-sectional and temporal variation could more rigorously identify causal drivers through fixed effects models, controlling for time-invariant country characteristics and global shocks.

7. References

1. Choi, T.-M., Yu, Y., & Au, K.-F. (2011). A hybrid SARIMA wavelet transform method for sales forecasting. Decision Support Systems, 51(1), 130–140. https://doi.org/10.1016/j.dss.2010.12.002
2. Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. International Journal of Information Management Data Insights, 2(1), 100058. https://doi.org/10.1016/j.jjimei.2022.100058
3. Hooijman, M. F., Martinez-De la Torre, A., Weiler, S., & Burden, A. M. (2022). Opioid sales and opioid-related poisonings in Switzerland: A descriptive population-based time-series analysis. The Lancet Regional Health - Europe, 20, 100437. https://doi.org/10.1016/j.lanepe.2022.100437
4. Li, H., Wu, Y. J., & Chen, Y. (2020). Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales. Journal of Computational and Applied Mathematics, 370, 112659. https://doi.org/10.1016/j.cam.2019.112659
5. Moolla, A., Holmes, J., Wilson, L., Brown, J., Kersbergen, I., & Stevely, A. (2025). Temporary and sustained changes in alcoholic and alcohol-free or low-alcohol drinks sales during January? A time series analysis of seasonal patterning in Great Britain. International Journal of Drug Policy, 145, 104939. https://doi.org/10.1016/j.drugpo.2025.104939
6. Romano, S., Galante, H., Figueira, D., Mendes, Z., & Rodrigues, A. T. (2021). Time-trend analysis of medicine sales and shortages during COVID-19 outbreak: Data from community pharmacies. Research in Social and Administrative Pharmacy, 17(1), 1876–1881. https://doi.org/10.1016/j.sapharm.2020.05.024
7. Van Calster, T., Baesens, B., & Lemahieu, W. (2017). ProfARIMA: A profit-driven order identification algorithm for ARIMA models in sales forecasting. Applied Soft Computing, 60, 775–785. https://doi.org/10.1016/j.asoc.2017.02.011
8. Van Ruitenbeek, R. E., Koole, G. M., & Bhulai, S. (2023). A hierarchical agglomerative clustering for product sales forecasting. Decision Analytics Journal, 8, 100318. https://doi.org/10.1016/j.dajour.2023.100318