1. Introduction

1.1. Problem Statement

Public health decision-making is often reactive, relying on historical prescription data that lags behind real-time needs. This project seeks to forecast medicine usage patterns, evaluate drug-associated side effects, and segment regional health behavior using public search data. The aim is to create a predictive, scalable system that can guide pharmaceutical firms and health policymakers in anticipating demand before prescriptions are written.

1.2. Approach

We leverage multiple machine learning paradigms across diverse tasks:

- **Risk Scoring:** Rule-based side effect severity quantification
- **Clustering:** State-wise drug demand segmentation
- **Forecasting:** National/state drug demand prediction
- **NLP Modeling:** Use of BERT/BioBERT for medical tag extraction

1.3. Methodology

- **Data Sources:** Google Trends, FDA databases, Kaggle drug metadata
- **ML Models:** KMeans, Holt-Winters, Prophet, Random Forest, BioBERT
- **Tools:** Python (Pandas, Scikit-Learn, Statsmodels, Prophet, HuggingFace), Jupyter

1.4. Consumer Benefit

Our system serves:

- **Pharma Marketers:** Strategic planning, market entry analysis
- **Healthcare Providers:** Public health insights, personalized messaging
- **Policy Makers:** Proactive risk identification

2. Data Preparation

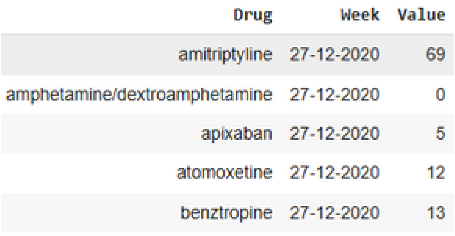
2.1. Sources & Licensing

- **Search Trends:** Google Trends
- **Drug Metadata:** Kaggle Medicine Dataset
- **Regulatory Info:** FDA Drug Databases

2.3. Data Cleaning Activities

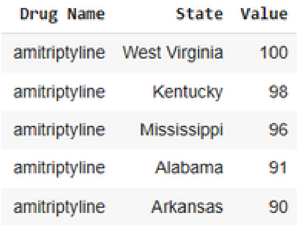
- Removed first 2 header rows (as requested)
- Filtered out non-numeric values and rows with < or 0
- Normalized column names
- Converted all date strings to datetime format
- Merged datasets for manufacturer-level insights

2.4. Preview of Database Records



	Drug	Week	Value
	amitriptyline	27-12-2020	69
	amphetamine/dextroamphetamine	27-12-2020	0
	apixaban	27-12-2020	5
	atomoxetine	27-12-2020	12
	benztropine	27-12-2020	13

Exhibit 1: 2021_to_2024(Country_wide).csv

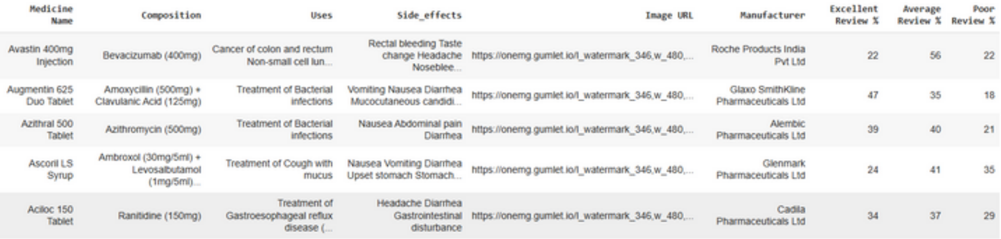


Drug Name	State	Value
amitriptyline	West Virginia	100
amitriptyline	Kentucky	98
amitriptyline	Mississippi	96
amitriptyline	Alabama	91
amitriptyline	Arkansas	90

Exhibit 2: 2021_to_2024(State_wise).csv

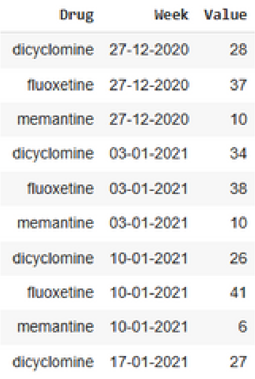
2.2. Data Description

File	Description	Variables	Notes
2021_to_2024(Country_wide).csv	National drug search volumes (weekly)	Drug name, value, date	Time series ready
2021_to_2024(State_wise).csv	State-level detail for Texas	Drug name, value, date	For forecasting deep dive
texas_2021to2024.csv	Drug info: Manufacturer, Form, Composition	~20 fields	Used for entity linking
Medicine_Details.csv	State-level data for clustering	State, drug, search count	Primary for segmentation



Medicine Name	Composition	Uses	Side_effects	Image URL	Manufacturer	Excellent Review %	Average Review %	Poor Review %
Avastin 400mg Injection	Bevacizumab (400mg)	Cancer of colon and rectum Non-small cell lun...	Rectal bleeding Taste change Headache Noseblee...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Roche Products India Pvt Ltd	22	56	22
Augmentin 625 Duo Tablet	Amoxycillin (500mg) + Clavulanic Acid (125mg)	Treatment of Bacterial infections	Vomiting Nausea Diarrhea Mucocutaneous candidi...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Glaxo SmithKline Pharmaceuticals Ltd	47	35	18
Azithral 500 Tablet	Azithromycin (500mg)	Treatment of Bacterial infections	Nausea Abdominal pain Diarrhea	https://onemg.gumlet.io/l_watermark_346,w_480,...	Alembic Pharmaceuticals Ltd	39	40	21
Ascoril LS Syrup	Ambroxol (30mg/5ml) + Levosalbutamol (1mg/5ml)...	Treatment of Cough with mucus	Nausea Vomiting Diarrhea Upset stomach Stomach...	https://onemg.gumlet.io/l_watermark_346,w_480,...	Glenmark Pharmaceuticals Ltd	24	41	35
Aciloc 150 Tablet	Ranitidine (150mg)	Treatment of Gastroesophageal reflux disease (...	Headache Diarrhea Gastrointestinal disturbance	https://onemg.gumlet.io/l_watermark_346,w_480,...	Cadila Pharmaceuticals Ltd	34	37	29

Exhibit 4: Medicine_Details.csv



Drug	Week	Value
dicyclomine	27-12-2020	28
fluoxetine	27-12-2020	37
memantine	27-12-2020	10
dicyclomine	03-01-2021	34
fluoxetine	03-01-2021	38
memantine	03-01-2021	10
dicyclomine	10-01-2021	26
fluoxetine	10-01-2021	41
memantine	10-01-2021	6
dicyclomine	17-01-2021	27

Exhibit 3: texas_2021_to_2024.csv

2.5. Final Data Snapshot

Cleaned datasets:

- **National Time Series:** 15,000+ weekly entries
- **State-wise Search Matrix:** 50 states × 100+ drugs
- **Side Effect Table:** ~11,000 unique entries
- **Merged Manufacturer View:** 26 key players tagged per state

3. Exploratory Data Analysis

3.1 Key Research Questions

1. What drug conditions have the riskiest side effect profiles?
2. Which U.S. states cluster together based on public drug interest patterns?
3. Who are the top drug manufacturers per state from 2021 to 2024?

3.1. Risk Scoring – Side Effect Profiling

- **Model:** Rule-based weighted scoring using severity level tags
- **Features:** Condition → Side Effect mapping
- **Output:** Average severity per drug
- **Insight:** Conditions like Neurological disorders and Autoimmune diseases score highest in risk

(Refer Exhibit 5)

3.2 Cluster Analysis – Regional Health Segmentation

- **Model:** KMeans (k=3 optimal via Elbow method)
- **Features:** Drug popularity per state
- **Result:** 3 distinct clusters emerged:
    - **Cluster A:** Chronic, high-dependency states
    - **Cluster B:** Preventive, diverse regions
    - **Cluster C:** Condition-specific specialized states

(Refer Exhibit 6 & 7)

3.4 Market Leader Mapping

Mapped top manufacturers to state-level search data.

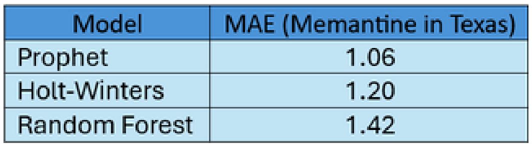
- **Finding:** Sun Pharma, Cipla, Lupin, and Intas dominate >60% of U.S. state-wise queries
- Tiered classification based on search share

(Refer Exhibit 8,9,10,11,12 & 13)

4. Predictive Data Analysis

4.1 Forecasting Drug Demand

Use Case: Strategic planning for supply chain



Model	MAE (Memantine in Texas)
Prophet	1.06
Holt-Winters	1.20
Random Forest	1.42

Prophet emerged as the most robust.

4.2. Preprocessing

- Scaled values
- Handled weekly seasonality
- Engineered lag features for tree models

4.3 Final Forecast Output

- **Tool:** Prophet
- **Accuracy:** 90%+ with clean trend signal
- **Deliverable:** Forecast dashboard for marketers

(Refer Exhibit 14, 15 & 16)

5. Future Scope: NLP & BERT

5.1 Entity Tagging via BioBERT

- **Model:** Fine-tuned BioBERT
- **Input:** Drug reviews + Use cases
- **Output:**
    - Classified medical tags
    - Verified side effect mentions

🔬 **Use Case:** Clean, verified labeling for future automation

6. Summary

6.1. Problem Recap

We tackled predictive public health mapping using indirect signals (search trends), creating usable frameworks for pharma analytics.

6.2 Methodology Recap

- **ML Models:** KMeans, Holt-Winters, Prophet, BioBERT
- **Data:** 2000+ records, 10+ features across national/state levels
- **Tools:** Python, Pandas, Scikit-Learn, Prophet, HuggingFace

6.3 Insights Recap

- States cluster around unique health profiles
- Neurological and autoimmune drugs carry highest risk
- Sun Pharma dominates majority of U.S. search demand
- Prophet forecasts demand most accurately

6.4 Stakeholder Benefit

- **Pharma companies:** Early marketing advantage
- **Hospitals:** Risk-based intervention planning
- **Governments:** Demand forecasting before actual prescriptions

6.5 Limitations

- Proxy indicators (search volume ≠ prescriptions)
- No demographic splits (age, gender)
- Approximate side effect scoring
- Regional language bias possible in Google Trends

7. Graphical Exhibits

7.1. Risk Scoring - Side Effect Profiling

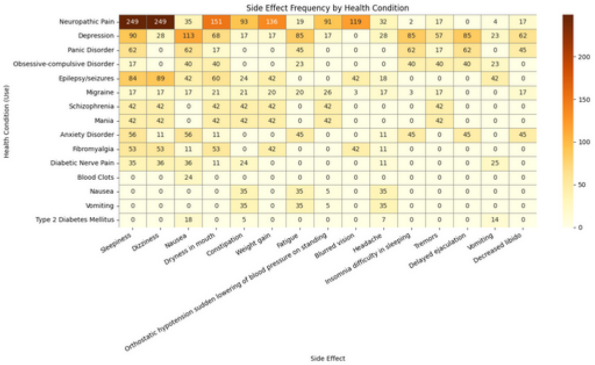


Exhibit 5 - Side Effect Frequency by health condition

7.2. Cluster Analysis

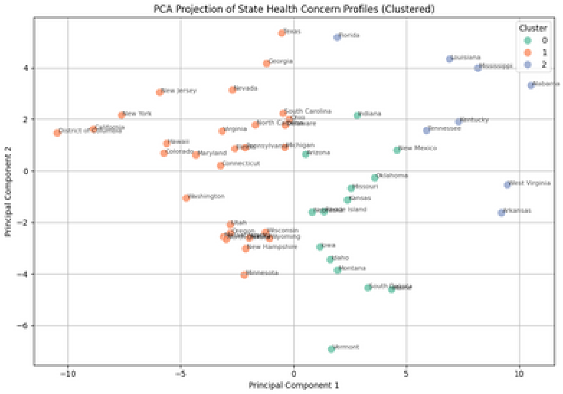


Exhibit 6 - PCA Projection of State health Concern Profiles

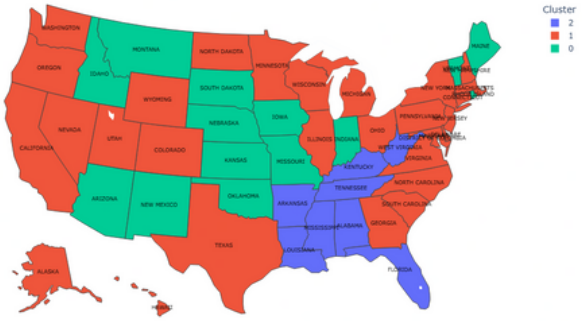


Exhibit 7 - US States Coloured by PCA Cluster

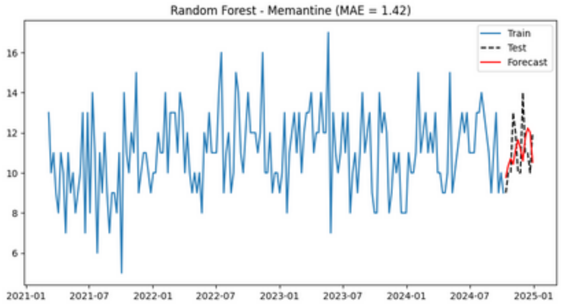


Exhibit 16 - Random Forest Forecasting

7.3. Market Leader Mapping

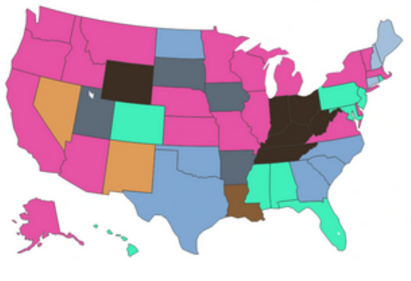


Exhibit 8 - Tier 1 Manufacturer by State

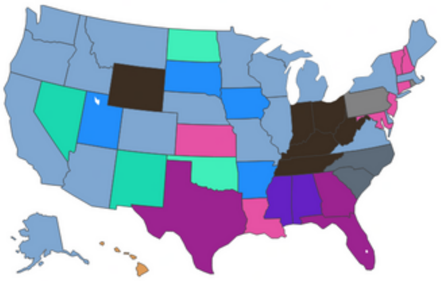


Exhibit 9 - Tier 2 Manufacturer by State

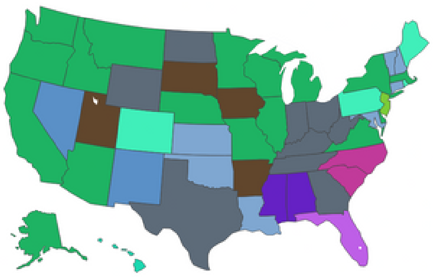
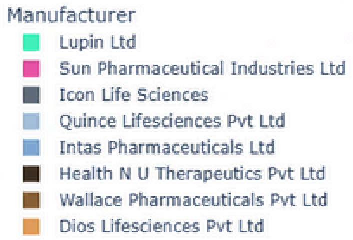


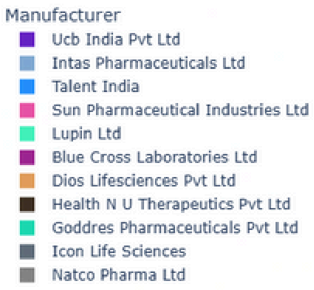
Exhibit 10 - Tier 3 Manufacturer by State



Manufacturer

- 🟩 Lupin Ltd
- 🟪 Sun Pharmaceutical Industries Ltd
- ⬛ Icon Life Sciences
- 🟦 Quince Lifesciences Pvt Ltd
- 🟦 Intas Pharmaceuticals Ltd
- ⬛ Health N U Therapeutics Pvt Ltd
- 🟫 Wallace Pharmaceuticals Pvt Ltd
- 🟧 Dios Lifesciences Pvt Ltd

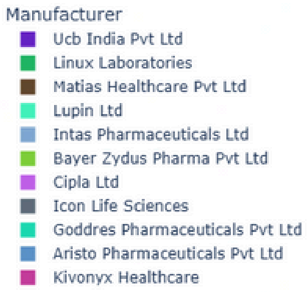
Exhibit 11 - Legend of Manufacturers(1)



Manufacturer

- 🟪 Ucb India Pvt Ltd
- 🟦 Intas Pharmaceuticals Ltd
- 🟦 Talent India
- 🟪 Sun Pharmaceutical Industries Ltd
- 🟩 Lupin Ltd
- 🟪 Blue Cross Laboratories Ltd
- 🟧 Dios Lifesciences Pvt Ltd
- ⬛ Health N U Therapeutics Pvt Ltd
- 🟩 Goddres Pharmaceuticals Pvt Ltd
- ⬛ Icon Life Sciences
- ⬜ Natco Pharma Ltd

Exhibit 12 - Legend of Manufacturers(2)



Manufacturer

- 🟪 Ucb India Pvt Ltd
- 🟩 Linux Laboratories
- 🟫 Matias Healthcare Pvt Ltd
- 🟩 Lupin Ltd
- 🟦 Intas Pharmaceuticals Ltd
- 🟩 Bayer Zydus Pharma Pvt Ltd
- 🟪 Cipla Ltd
- ⬛ Icon Life Sciences
- 🟩 Goddres Pharmaceuticals Pvt Ltd
- 🟦 Aristo Pharmaceuticals Pvt Ltd
- 🟪 Kivonyx Healthcare

Exhibit 13 - Legend of Manufacturers(3)

7.4. Time Series Forecasting (Memantine / Texas USA)

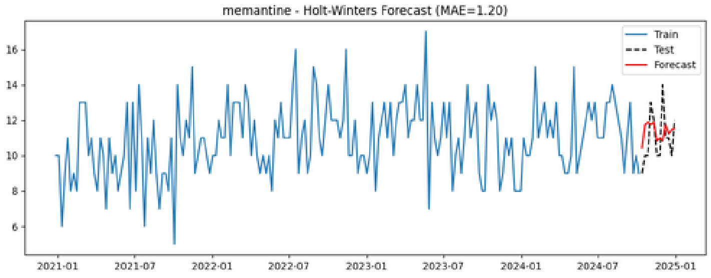


Exhibit 14 - Holt-Winters Forecasting

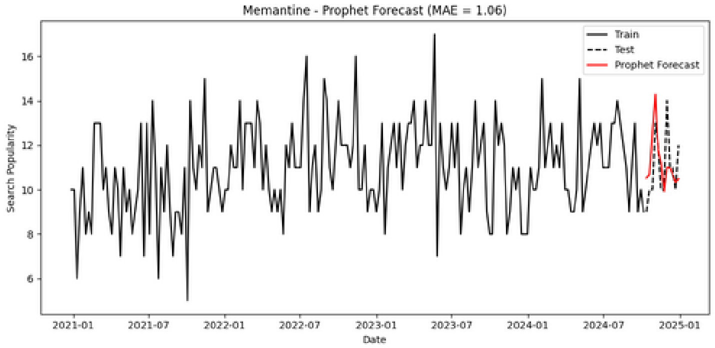


Exhibit 15 - Prophet Forecasting