,

INNOPIPHANY DRUG SPENDING FORECAST

Authors: Rose Chittilappilly, Jimin Heo, Sayalee Joshi, Jasmine Son, Minal Walvekar, Hazel Yu
Collaborators: Cody Dunn (cody.dunn@innopiphany.com), Shawn DeLuz (shawn.deluz@innopiphany.com)
Instructors: Faisal Nawab, Brian G Vegetabile
TAs: Jizhi Zhang, Jinhwa Kim

1 EXECUTIVE SUMMARY

- **Introduction:** This project aims to provide Innopiphany with a forecasting tool that aids in early-stage pharmaceutical companies about anticipated government spending for their upcoming drugs.
- **Project Goal:** Our goal with this project is to forecast the expected government spending for an emerging drug based on existing analogs for the indicated disease of choice and route of administration and integrate this into a user interface.
- **Problem Need:** Early stage pharmaceutical companies do not have enough resources to create accurate forecasting methods for their upcoming drugs.
- **Value proposition:** The user interface takes in a disease name as an input and optionally takes in preferred routes of administration to provide information about and forecast drug prices of the disease state of choice. The tool compares top performing analogs for a specified disease and forecasts changes in spending based on the years after approval for new drugs in the same disease space.
- **Product Overview:** The interface displays graphs and key metrics for total spending, average spending per beneficiary, total number of beneficiaries, predetermined growth forecasting, and curve-fitting growth forecasting.
- **Resources:** We utilize the openFDA API to call the drugs@FDA and product labeling datasets for the analog information and intertwine it with the Medicare Part D spending data to create our dataset. We work primarily with Python in Visual Studio Code to write our script, create a GitHub repository, and use Streamlit for the user interface.
- **Conclusion:** This forecasting tool should help early-stage pharmaceutical companies in understanding the market trends of drugs for a specified disease space and give insight into potential Medicare spending for their new drug products.

2 OVERVIEW & PROJECT GOALS

When pharmaceutical companies launch new drugs, it is important for them to predict potential profits to make investment decisions. However, early-stage companies do not have enough data to create accurate forecasts and often look at the number of patients for a disease to predict government spending. This project aims to forecast overall market trends based on the performance of existing analog drugs in the market and integrate this into a tool that can perform a semi-automated analysis of the analogs.

Ultimately, we aim to develop a model to aid early-stage pharmaceutical companies in predicting the potential financial performance of their newly developed drug and delivering it through the user interface.

3 BACKGROUND INFORMATION

Innopiphany is a life science consulting company focused on analytics, modeling, and forecasting. With Innopiphany, we created a model that forecasts the Medicare spending of drugs to aid early-stage pharmaceutical companies in getting investments for developing new drugs. The model compares different analogs, which are different drugs used to treat the same condition, and their spending to inform the forecasting model to project expected Medicare spending per disease.

Before the launch of a new drug, large pharmaceutical companies will conduct a rigorous forecast of their expected financial performance to secure investors and to accurately price their drugs to maximize profit. Smaller, early-stage pharmaceutical companies do not have the same resources to make accurate predictions.

As mentioned previously, early-stage companies often use an epidemiology-based approach, where they forecast sales based on the number of patients with a disease. With this method, the patient population is the primary driver of demand. While it can give insight into the potential market size, this method does not capture real-world market dynamics that can affect spending patterns. However, an analog-based method forecasts based on the government spending of similar drugs. This means that historical government spending is the primary driver of demand. This is more accurate for forecasting spending because it provides a real-world context and should capture market dynamics.

Therefore, we used an analog-based approach to create spending forecasts that more accurately reflect market trends within each disease space.

3.1 Datasets

The datasets used in this model are the Medicare Part D Spending data and the drugs@FDA and product labeling data from the openFDA API.

- **The Medicare Part D data** contains information regarding the spending of drugs prescribed nationally to Medicare beneficiaries enrolled in Part D, a program that helps pay for prescription drugs and vaccines. It shows information such as the total spending, which shows how much Medicare spent in the entire year in total for each drug. The dataset also shows total beneficiaries and the average spending per beneficiary. This shows how many individuals Medicare is spending money on as well as the average amount of medical spending each person requires for the specific drug. This dataset only has data from the years 2018-2022.
- **OpenFDA** is an API that provides public FDA data about drugs, devices, and foods in JSON format. Drugs@FDA specifically provides information about drug names, active ingredients, dosage forms, routes of administration, latest and previously approved labels, and regulatory information. The Product Labeling data contains information on the FDA drug label, including the brand name, generic name, and indications and usage.

We combined these datasets to create a more comprehensive view of the drug's financial and product information. We joined the datasets on the brand name and the generic name of the drugs which was present in both of the datasets from the openFDA API as well as in the Medicare Part D data.

A feature that was crucial for this forecasting was the *"indication_and_usage"* field of the Product Labeling data. This provides us with the diseases that a drug is indicated to treat. This field is used when the user specifies which diseases they are interested in forecasting drug prices for. When the user types in the disease name of choice into the interface, the API will call drugs indicated for the specific disease for a more accurate forecasting.

Another important feature of the Product Labeling data was the "route," or the route of administration. The route specifies how the drug is administered to the user, including but not limited to orally, subcutaneously, and intravenously. This is essential because the route of administration can largely affect the cost of the drug. Including this feature gives the user the option to filter for specific routes of administration depending on their forecasting needs.

The *"submission_status_date"* was also essential for the scope of this project. As previously mentioned, the financial data only spans from 2018-2022, which is a relatively small time frame. This does not show the entire picture of how a drug

performs throughout its time in the market if it was released much earlier than 2018. We used the *submission_status_date* as an approximate alternative to the drug's approval date to understand which portion of the drug's market lifespan the financial data corresponds to, whether it is in the early stage or several years after approval. This can give insights on how a drug's spending changes across the years after approval.

3.2 Limitations

Some limitations of this project are in regard to the datasets used. To inform the model of drug spending, we used the Medicare Part D data. The drawback of this dataset is that Medicare primarily serves the population that is ages 65 years or older, which restricts the applicability of our findings to populations outside of Medicare, particularly younger demographics.

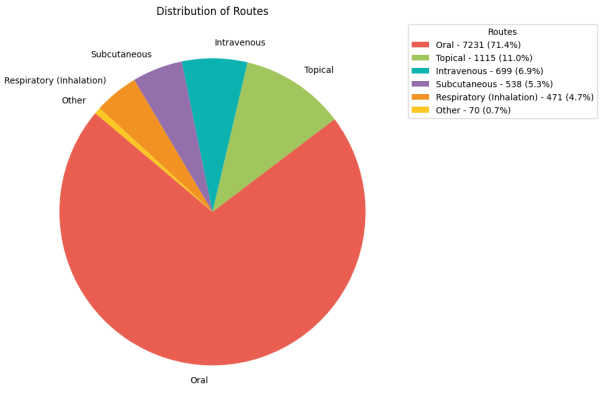
Medical data is often highly proprietary, protected by stringent privacy regulations, and expensive to acquire, particularly for startups and smaller companies. Due to this, we were limited in using publicly available data, but we expect that early stage pharmaceutical companies would face similar limitations.

3.3 EDA

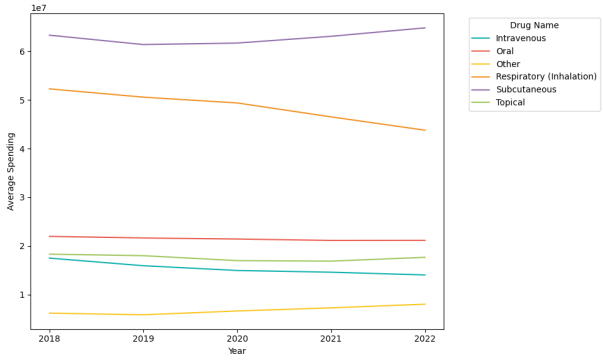
- Initially, we analyzed the top five drugs with the highest total spending within each disease area to ensure the API calls were accurately retrieving the relevant data. To support this validation, we used a reference list of drugs expected to appear, given to us by domain experts, based on the specified disease area. Once we confirmed the accuracy of the API results, we proceeded with a deeper analysis.

  The final user interface incorporates this analysis, providing insights into the top five drugs for any inputted disease area and their performance from 2018 to 2022. This includes data on total spending, average spending per beneficiary and the number of beneficiaries over the five-year period. The main reason of focusing on the top 5 drugs and not the worst performing drugs as the drugs with lower spending are often been in the market for a long time, with the patents expired. This would not be a good comparison for the newly launched drugs.

- We explored the route of administration for drugs, which is the way in which the drug is taken in by the consumer. This is an important factor during forecasting because spending might depend on the type of drug that is being developed. Figure 1a shows the distribution of the routes of administration for all the drugs in the Medicare spending dataset. Majority of the drugs in the dataset are taken orally (71.4%), but this does not mean that Medicare spends the most on oral drugs.



(a) Spend per Route of Administration



(b) Spend per Route of Administration

The average spending per route of administration for all the drugs in the Medicare dataset is shown in Figure 1b indicates that Medicare spending per drug is highest for subcutaneous drugs (which only make up 5.3% of all the drugs), while the spending per drug is third highest for oral drugs. This could suggest that subcutaneous drugs are more expensive to make, and therefore must be priced higher. It could also suggest that the government might be

more willing to spend more money on subcutaneous drugs. Although the exact reason for this trend needs further investigation, we can see that there are clear differences in Medicare spending on drugs, depending on the route of administration. This led us to include the route of administration in our model.

4 METHODS

4.1 Forecast Method 1: Predefined Growth Rates

The first method of forecasting uses predefined growth rates provided by the Innopiphany team based on their data and expertise on how much growth in spending is observed. The growth rates are:

Year	Growth Rate
Year 1	11%
Year 2	31%
Year 3	58%
Year 4	76%
Year 5	89%
Year 6-10	100%

Table 1: Predefined Growth Rates

For each disease area, two benchmark drugs were selected, one with the highest Medicare spending in 2022 and another with the highest growth rate in spending between 2018 and 2022. These benchmark drugs were chosen to represent successful drugs within the disease space, for newer pharmaceutical companies to attract potential investments. The best-case scenario for the performance of drugs was chosen to show investors the current market trend. The worst-case scenario was not projected because spending would be zero, which is not very informative. The predefined spending growth rates were used to project future expenditures for the next 10 years (from 2024 to 2033) using the Medicare spending in 2022 of successful drugs as the highest point.

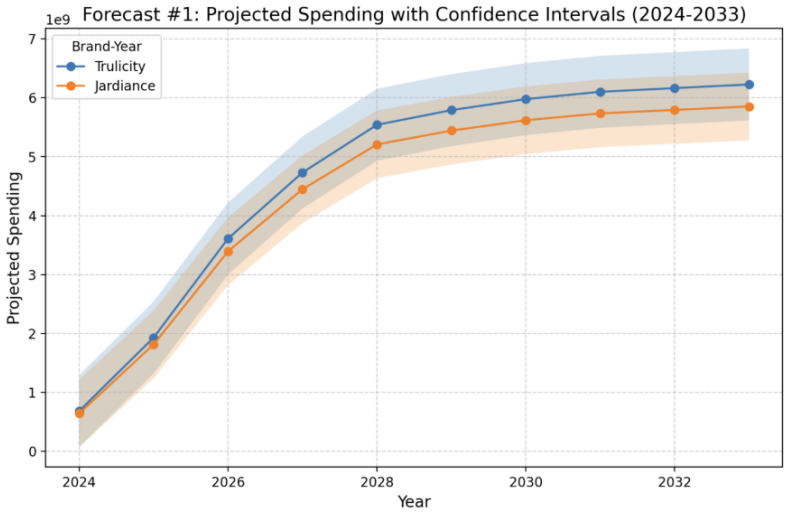


Figure 2: Forecast 1 : Type 2 Diabetes Mellitus

To estimate uncertainty around these projections, a fixed standard error, calculated as 5% of the 2022 spending, was assumed for simplicity. Using a normal distribution, the 95% confidence intervals were determined by scaling the standard error with

the critical value of 1.96 and applying it to the projected spending estimates. This approach provided upper and lower bounds for the anticipated spending trajectory.

Figure 2 shows the projected spending and confidence interval for a new drug to treat type 2 diabetes mellitus based on the spending for Trulicity, the drug with the highest spending in 2022 and for Jardiance, the drug with the highest slope from 2018-2022.

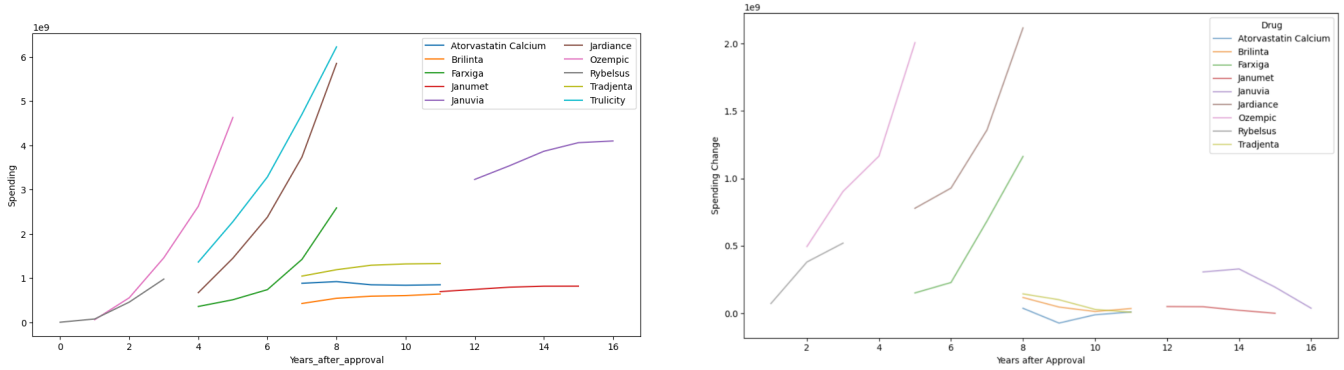
4.2 Forecast Method 2: Curve Fitting

The second method forecasts the spending change of newly launched drugs by using the number of years since the approval date for analogs in the same disease space. Since we need to analyze the growth change within their life cycle, we focused on drugs that are performing well financially, which are more likely be in the early stages of their life cycle. Similar to Forecast 1, this forecasting method focuses on successful drugs in the market within the same disease space. Therefore, the top 10 analogs were chosen based on the Medicare spending in 2022, since this would give a more realistic market overview. This approach involves a five-step process.

Step 1: The years since approval were calculated by subtracting the approval year from the years 2018-2022, instead of using the static years.

The graph in figure 3a) shows the spending based on years after approval of each drug for Type 2 diabetes mellitus. Each line begins in 2018, which marks the starting point of our data. For example, the pink line representing Ozempic starts at year 1, indicating that the drug was approved in 2017. These graphs allow us to observe different trends between relatively new drugs and those that have been on the market for a longer time. While newer drugs show exponential growth, older drugs tend to plateau after a certain period of time. Additionally, we observed two distinct groups of relatively older drugs: Januvia (purple) and Janumet (red). While their slopes are somewhat similar, the total spending shows significant differences.

Step 2: The change in spending was calculated based on these years.
Instead of focusing on total spending, the change in spending was prioritized, as our primary interest is the growth rates over the years following drug approval. (Fig 3b)



(a) Spending over Years After Approval

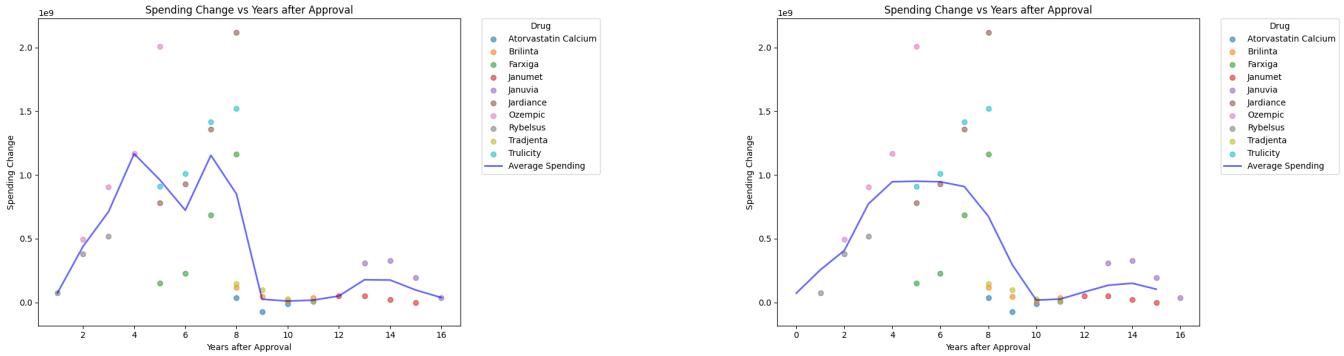
(b) Spending Changes over Years After Approval

Figure 3: Spending (Years of Approval)

Step 3: The results were averaged for each year after approval and a moving average across three years was applied to smooth the curve. (Fig 4)

Step 4: The cumulative sum of the spending change was calculated to get the projected spending.

Step 5: The standard deviation of spending change for each year after approval was calculated to generate error bars. (Fig 5)



(a) Average Spending over Years After Approval

(b) Moving Average Spending over Years After Approval

Figure 4: Curve Smoothening

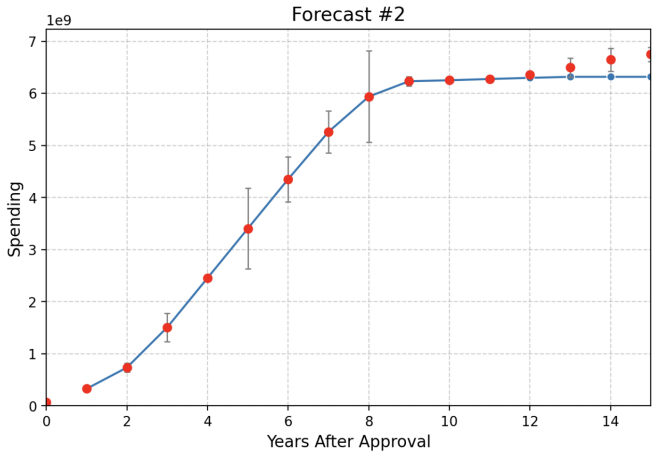
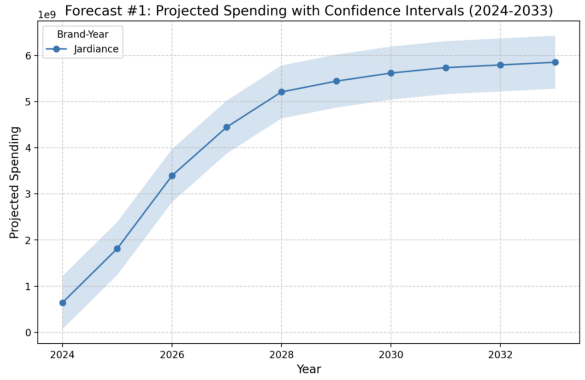


Figure 5: Forecast Method 2 with Error Bars

4.3 Effects of Route of Administration:

As previously highlighted, the route of administration plays a critical role in predicting total spending. For Type 2 diabetes mellitus, while oral drugs are more widely distributed, the government spends significantly less on them. In contrast, subcutaneous drugs, despite their lower distribution, incur higher spending from the government.

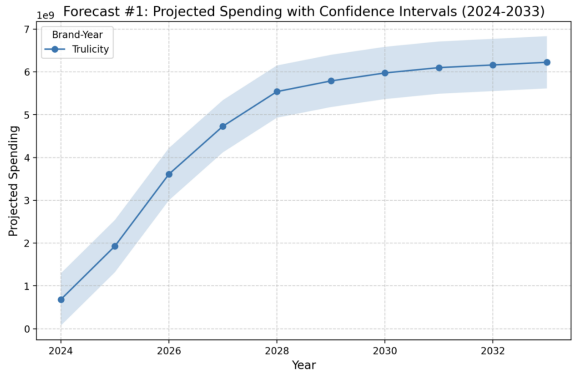
To illustrate this, we analyzed spending across various routes of administration, focusing specifically on oral and subcutaneous drugs. Changing the route of administration for forecast 1 changes the benchmark drug we use to forecast. We can observe here that even in the top performing drugs for Type 2 diabetes mellitus, the route of administration has some effect on the projected spending. As seen in Figure 6a, the peak projected Medicare revenue for oral drugs is approximately $5.8 billion, whereas the peak projected Medicare revenue for subcutaneous drugs (Fig 6b) is about $6.2 billion.



Key Metrics

- **Drug Name:** Jardiance
- **Peak Projected Revenue:** $5,851,720,782.00
- **Year of Peak Projected Revenue:** 2033
- **Total Projected Revenue:** $43,946,423,072.82
- **Average Revenue per Year:** $4,394,642,307.28
- **Route of Administration:** Oral

(a) Diabetes - Oral drugs



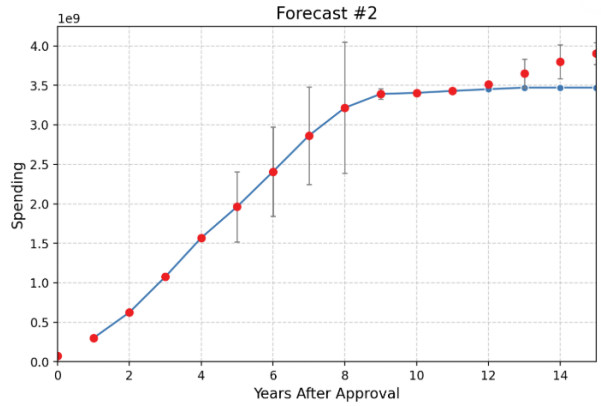
Key Metrics

- **Drug Name:** Trulicity
- **Peak Projected Revenue:** $6,225,291,667.60
- **Year of Peak Projected Revenue:** 2033
- **Total Projected Revenue:** $46,751,940,423.68
- **Average Revenue per Year:** $4,675,194,042.37
- **Route of Administration:** Subcutaneous

(b) Diabetes - Oral drugs

Figure 6: Forecast 1 : Routes of Administration (Type 2 Diabetes Mellitus)

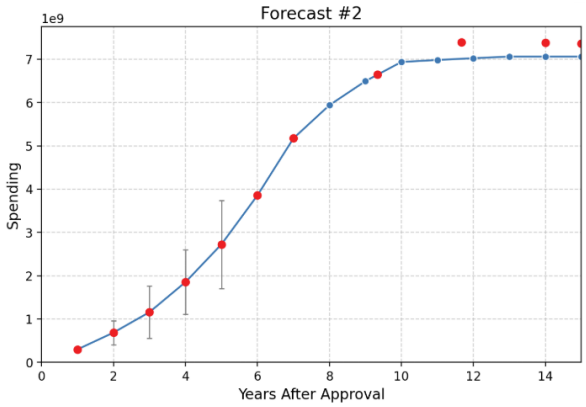
The peak projected Medicare revenue also changes in forecast 2 depending on the route of administration. As seen in Figure 7a, the peak projected Medicare revenue for oral drugs is approximately $3.5 billion, whereas the peak projected Medicare revenue for subcutaneous drugs (Fig 7b) is about $7 billion.



Key Metrics

- **Peak Projected Revenue:** $3,471,524,744.29
- **Year of Peak Projected Revenue:** 13
- **Total Projected Revenue:** $55,470,798,229.20
- **Average Revenue per Year:** $2,773,539,911.46
- **Route of Administration:** Oral

(a) Diabetes - Oral drugs



Key Metrics

- **Peak Projected Revenue:** $7,063,212,034.86
- **Year of Peak Projected Revenue:** 13
- **Total Projected Revenue:** $105,651,030,820.09
- **Average Revenue per Year:** $5,282,551,541.00
- **Route of Administration:** Subcutaneous

(b) Diabetes - Oral drugs

Figure 7: Forecast 2 : Routes of Administration (Type 2 Diabetes Mellitus)

4.4 Alternative Approaches

For the curve-fitting forecast model, using a moving median was considered instead of a moving average to smooth out the curve for change in spending. After testing using the median of 2-5 years, using the median across 3 years was the best. However, using a moving median did not smooth out the curve as much as a moving average did, so we ultimately chose to use a moving average. (Fig 8)

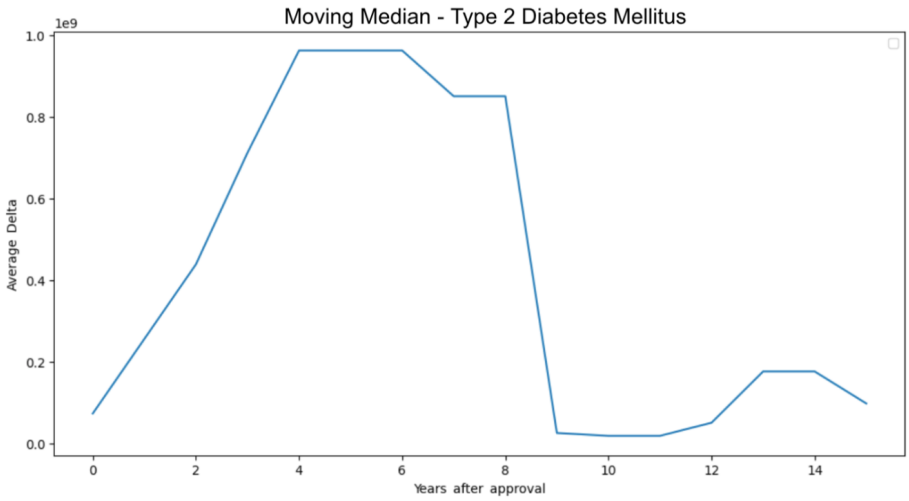


Figure 8: Forecast 2 : Moving Median

5 USER INTERFACE

When launching the interface, the user can input the desired indication. The openFDA labels do not have consistent names for some indications, so the code will generate some variations of the indication name, including spaces, dashes, and acronyms (eg. nonvalvular atrial fibrillation, non-valvular atrial fibrillation, non valvular atrial fibrillation, NVAF). The user can optionally select the desired route of administration for the analysis which include oral, topical, intravenous, subcutaneous, respiratory (inhalation), and other. Selecting one or more of these will filter the results to only include drugs with those routes of administration. If no routes are selected, it will return the results from all drugs.

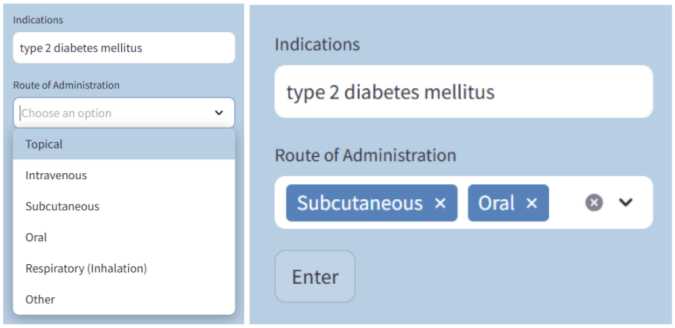


Figure 9: User Interface Input

The first tab of the dashboard displays information about the total spending for the top five drugs between 2018-2022 (Fig. 10). Similarly, the second tab displays information on the average spending per beneficiary for the top five drugs. The third

tab displays information on the total number of beneficiaries for the top five drugs. These three tabs also display key metrics, including the names of the top drugs, the drug with the highest spending, the drug with the highest growth, and the route of administration. The fourth tab displays the forecast based on the first forecasting method, which used a predefined growth rate. The fifth tab displays the forecast based on the second forecasting method along with the key metrics. The last two tabs also display key metrics, including the peak projected revenue, year of peak projected revenue, total projected revenue, average revenue per year, and the route of administration.

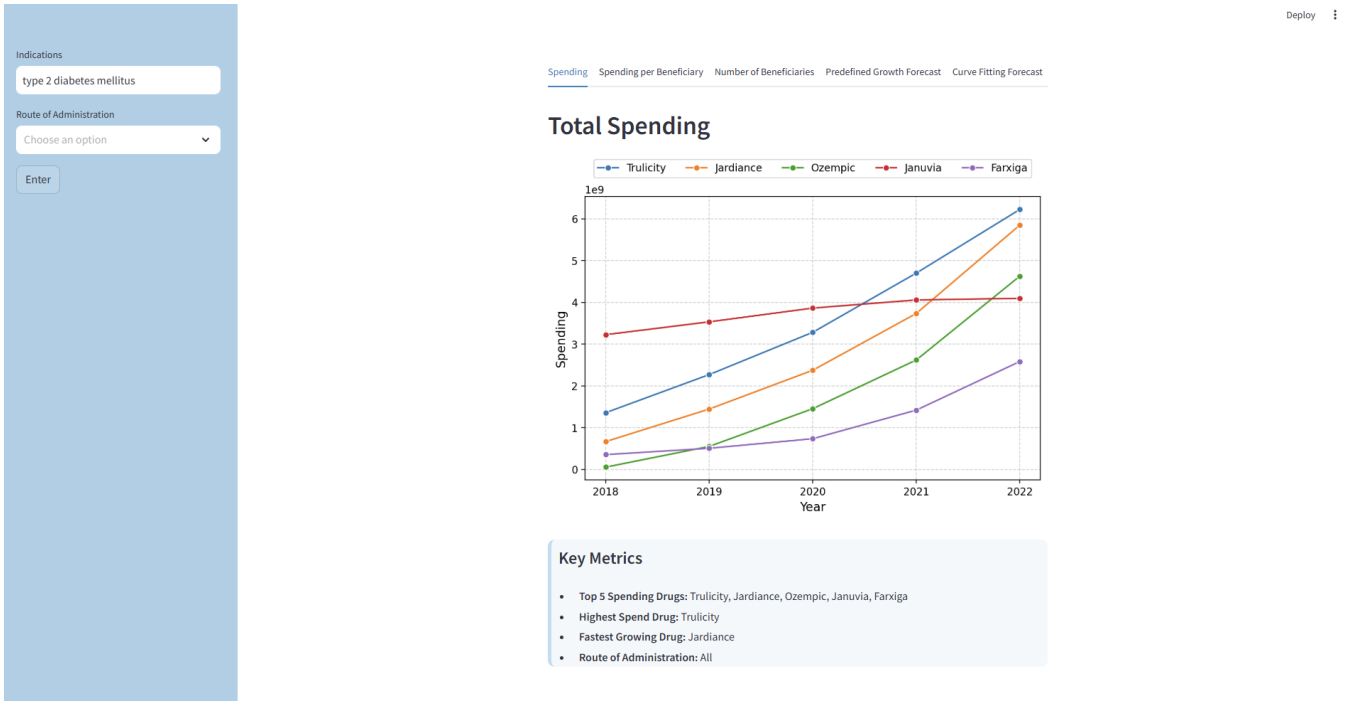


Figure 10: User Interface

Refer to Appendix A3, for additional details on the user interface.

6 VALIDATION

Our forecasting model provides potential investors with an estimate of the financial opportunity available in a specific disease area for a newly launched drug. The primary goal is to project the drug's financial potential 10 years post-launch, using current government spending in the targeted disease area.

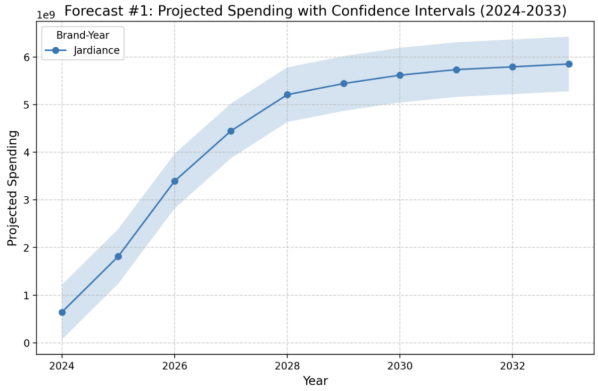
There were a few challenges in validation of our model. The available financial data spans only five years, offering insufficient historical context for deep learning or complex machine learning approaches. Drugs at different stages of their life cycle exhibit diverse spending patterns, which complicates model validation and introduces outlier behavior. Due to the limitations of available data and the absence of traditional validation techniques, the validation of our model relies heavily on domain knowledge and expert insights.

Forecast Method One leverages numbers provided by the Innopiphany team, who base these values on extensive domain expertise and historical observations. This method serves as a benchmark for evaluating our model's performance.

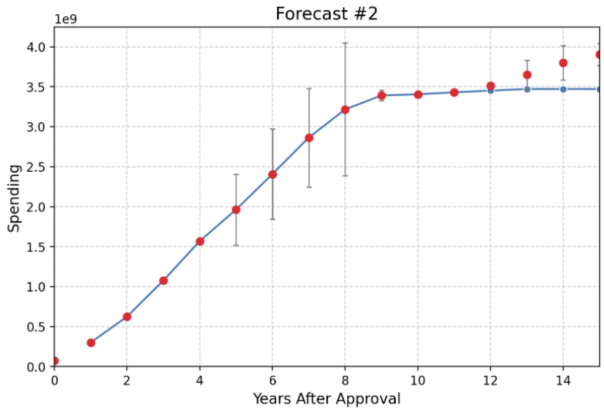
One of the key methods for assessing our model is verifying its ability to replicate observed trends in drug life cycle spending. These trends include periods of rapid increases in spending during early post-launch years and plateaus in spending as the drug matures in the market. Our model demonstrates its validity by aligning with these expected trends, providing confidence that it captures the underlying dynamics of drug spending over time. Unlike the predefined models, which focus solely on the

top-performing drugs within a disease area, our model incorporates a broader dataset that includes a wider range of drugs. This approach enhances its applicability by providing more accurate projections for average-performing drugs, which are often of greater interest to stakeholders seeking realistic market assessments. To quantify the uncertainty in the predictions, we included error bars in our model's outputs.

As seen in figure 11, in both forecast models, spending shows rapid growth in the first six years after a drug's launch and then tapers down. Forecast 1, shows peak spending around $5.8 billion, whereas Forecast 2 shows peak spending around $3.5 billion. This difference confirms that Forecast 1 skews toward high-performing drugs, which could be helpful for attracting potential investors, while Forecast 2 sets more realistic expectations.



(a) Forecast 1



(b) Forecast 2

Figure 11: Validation: Comparison between Trends

7 CONCLUSION & NEXT STEPS

The implemented model forecasts overall market trends based on the performance of existing analog drugs using historical spending data and integrates this into a user interface that can perform a semi-automated analysis of the analogs. This tool aids early-stage pharmaceutical companies looking to develop and launch new drugs in predicting the potential financial performance.

Currently, API calls take approximately 3 to 30 seconds to retrieve all drug names, depending on the number of drugs available for the disease and the system's computation performance. Once the drug names are loaded, the application simply sorts them by the route of administration to do further analysis. Thus, after consulting Innopihany, we determined that caching is not necessary for the current model.

We can extend this model to cover younger populations by including data from Medicaid and insurance companies since Medicare is generally for people aged 65 years and older. Including a wider demographic will increase the applicability of our model. Furthermore, future modifications can include web-scraping to gather data about other drugs launches and implement those market trends into the forecast. Currently, our model only takes in retrospective data on drug spending, so future work can include the launch date of competitors.

If given access to more data, more features could be included in our model by using a weighted average for spending change. Some features to potentially include are drug efficacy, number of drugs in the market, and the number of indications a drug can treat. Our model only gives an overall forecast for drugs in a specified disease space. While it does include route of administration, using more features in the future could help tailor the forecast to a client's specific drug.

8 APPENDICES

A1. **Team Member Contribution Statements:**

The project involves developing a code to interact with the openFDA API, performing exploratory data analysis (EDA) with insightful visualizations, and constructing two forecasting models for prediction. Additionally, the project includes designing and building a user interface (UI), followed by integrating the code into the final UI for optimal functionality and user experience. Everyone collaborated on each part equally. Some of the more detailed roles are outlined below:

Rose Chittilappilly: Wrote a function to standardize disease names for the API call, converted the Forecast methods into functions to integrate with the User Interface. Integrated functions to display the graphs onto the user interface. Added error bars to validate Forecast 2. Worked on debugging and issue fixes throughout the project. Contributed in report writing.

Jimin Heo: Wrote a function to make API calls to get the drugs for a particular disease. Worked on Forecast method 2 to get average spending changes per number of years after approval. Integrated the code into the user interface. Worked on debugging and issue fixes throughout the project. Contributed in report writing.

Sayalee Joshi: Wrote function to get Forecast 1 graph. Worked on extrapolation techniques in Forecast 2 for missing data. Wrote a function to fix duplicate drug brand names in Forecast 2 to get an accurate forecast. Integrated functions to get graphs and key metrics in the user interface. Contributed in report writing.

Jasmine Son: Conducted exploratory data analysis to get insights from the data to create visualizations. Wrote functions to implement the route of administration of drugs into the Forecast methods. Worked on developing the user interface, extracted and displayed key metrics in the UI. Contributed in report writing.

Minal Walvekar: Wrote a function to standardize disease names for the API call. Conducted EDA to visualize the trends. Implemented route of administration of drugs into the Forecast model. Added confidence intervals in Forecast 1 to validate the results. Contributed in report writing.

Hazel Yu: Wrote code to input disease name in the drug API. Implemented a function to get drug approval dates for Forecast 2. Implemented moving averages to smoothen the curve fitting forecast. Developed the initial framework for the user interface and integrated the code into the UI. Contributed in report writing.

A2. **Additional EDA:**

Patent Expiration

A key factor influencing trends in drug spending is the expiration of patents on brand-name drugs. These drugs tend to be significantly more expensive because they are protected by patents, which prevent other companies from producing generic versions during the patent period. This lack of competition allows the original manufacturer to set higher prices. However, once the patent expires, generic versions of the drug can be introduced to the market. Generics, while bio equivalent to the brand-name drug, are sold at a fraction of the cost, resulting in a decrease in prices due to increased market competition.

Thus, we can expect brand-name drugs to dominate spending trends during the period when their patents are still active. For example, the patent for Biktarvy is set to expire in 2033, while Tivicay's generic launch is anticipated in June 2030. This is reflected in the graphs, where these drugs show an upward trend in spending as they remain patent-protected. (Fig.12)

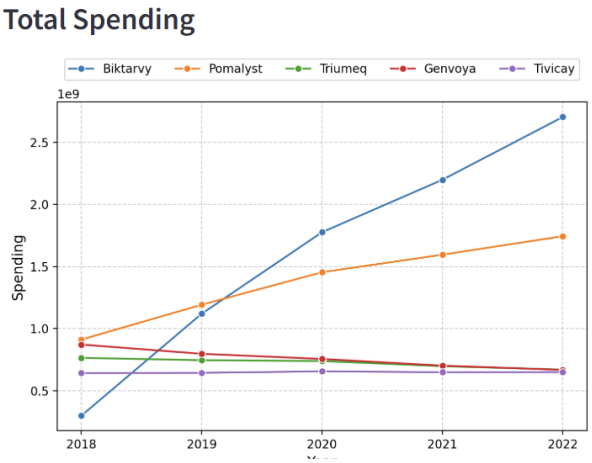
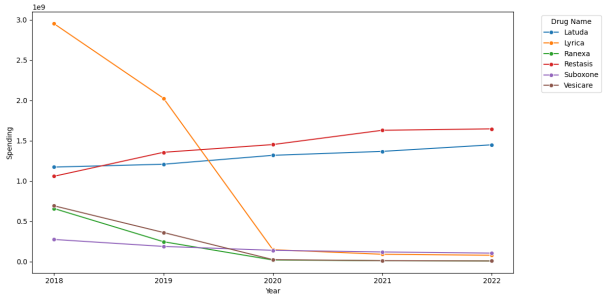
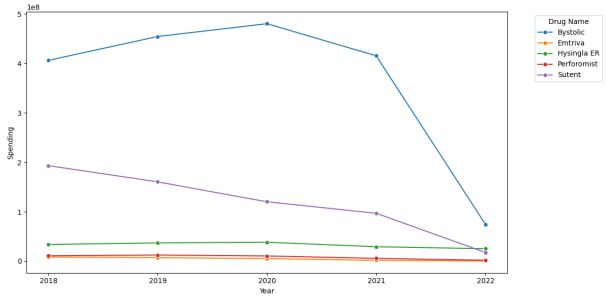


Figure 12: Top 5 highest spending drugs for HIV

As seen in the graphs, there is a notable drop in spending after the patent expires. (Fig.13) We explored whether the drop in spending that typically follows patent expiration could be incorporated into our forecast models. However, this factor was not included in our final forecast model, as most patents are not expected to expire for at least 20 years, and our focus was on the first 10 years post-launch when the drug is still under patent protection. Therefore, the effect of patent expiration on spending was deemed less relevant for our analysis.



(a) Drugs with Patents that expired in 2019



(b) Drugs with Patents that expired in 2021

Figure 13: Patent Expiration

A3. **Additional System or Methods Details:**

- User Interface Demo Video
- Technical Document

A4. **Links to Datasets:**

- Medicare Part D Spending Data
- openFDA