Project Description Document

Author: Manvi Semitha
Project title: Quantifying The Rate at which Transformative Drugs are Developed
Class: DS2065 Data Science- Theory and Practice

Objective/Question:

Most treatments are incremental, a few are "transformative"- far better than anything else available for the condition they treat. With my research my aim is to quantify how often transformative treatments come around and how? I want to look into the discovery and development years of the transformative drug and see what factored into their discovery and whether there is any correlation between certain factors and the drug's creation.

The reason why I chose this question stems from my future aspirations as someone who wants to delve into Biotech for further studies after undergrad. Originally I thought it would be interesting to look into the rates of drug development over the years. My question quickly became more specific to transformative or breakthrough drugs as I went on with my research and found the field quite fascinating especially as I happened upon the 2017 creation/discovery of the CAR T-cell therapy and the whopping price tag associated with the treatment due its status as "transformative." I quickly realised that this field is something of great financial interest to Biotech companies with a general trend of R&D funding steadily increasing over the years.

I selected the FDA Breakthrough Drugs list as my dataset on the FDA's official website to explore the rates at which these drug were being approved. I also conducted individual research on drugs I found interesting or fascinating for example, Aduhelm- the first drug approved for the treatment of Alzheimer's and the controversy surrounding it.

Legal and Ethical Implications

Considering the legal implications of every aspect of my project, I ensured that all data was handled responsibly. I did this by only using publicly available datasets and conducting thorough research on drugs on the list that were considered controversial. I realised that there were some ethical limitations to this analysis as well, as there is an aspect of:
Data Bias: Reliance on FDA data excludes non-U.S. approvals, potentially overlooking global innovations and skewing results toward Western medical priorities.

Commercial Influence: Pharmaceutical companies may prioritize profit-driven "transformative" drugs limiting affordability and accessibility as the price tag of breakthrough treatments often raises questions about equity in healthcare, with concerns about whether these treatments will be accessible to the patients who need them most.

Use and Analysis

The dataset I used was:

- FDA Breakthrough Drugs List 2012-Current: https://www.fda.gov/media/95302/download

I found the dataset while looking for an accessible and dependable drugs list online and decided to use it for my study. However, the set required significant efforts for extraction as it only came as a .pdf file.

I prepared my dataset as follows:

1. Data was extracted from the FDA Breakthrough Drugs PDF list with a tool called Tabula.
2. The resulting CSV file was coded and formatted using DeepSeek.
3. Using an online CSV converter the code was then exported into a new CSV file.
4. The CSV file was then imported into excel where it was delimited by ","
5. The data was cleaned- unavailable data was removed, and cells were formatted into text and number cells.
6. The datasets were in table CSV format where additional precautionary methods like .dropna() was used to better prepare the data.
7. The data was visualised through tools such as seaborn and matplotlib.

I then created line graphs, heatmaps, histograms, stacked bar plots and linear regression models to analyse my data on the basis of "Top 10 companies by approvals per year," "Number of drug approvals per year," "The rate of first-in-class drug approvals per year" and "Predicting Transformative Drug Approvals for the coming years."

The analysis of FDA-approved drugs from 2012 to 2022 indicates a slight upward trend in the approval of breakthrough drugs, despite year-to-year variability. A predictive linear regression models, trained on historical data up to 2020, forecasted a steady increase in the annual approval of transformative and FIC drugs through 2030. For instance, the model predicts an increase from approximately 24 transformative drugs in 2021 to around 46 by 2030.

Moreover, there is an uptick in transformative drugs approved in 2017-2018 due to discovery of CAR-T cell therapy by Novartis. Genentech- a research driven biotech company has been consistently contributing to breakthrough drugs since 2012. There is a peak in 2020 due to the FDA expediting approval processes attributed to Covid-19.

This analysis seems to indicate that the rate at which transformative drugs are developed has been steadily rising with the number of drugs developed predicted to almost double over this decade. New technological developments, and external pressures such as covid-19 seem to be driving innovation as these situations expedite FDA approval processes.

Learning Outcomes

In the process of finding a dataset, I explored several databases but realised that there is a seeming lack of publicly available data for drug approvals and drug lifecycles. While they do exist, they are more often than not either hidden behind a paywall with a hefty fee or are extremely complicated to sort and make sense of due to disorganisation and a paralysis in terms of the amount of information scattered across different biotech company websites. Regulatory websites like the FDA or scientific journals were my best bet however I came across the same problems as journals required you to sign-up and pay a fee or regulatory sites did not have information available as downloadable .csv files. The extraction process because of this was gruelling and took several days to figure out and yet I was not able to decrypt data beyond 2020 due to processing limitations.

Due to my previous experiences with data manipulation, I was familiar with the task and was comfortable cleaning up data but I also further learned about implementing linear regression models for predicting general trends,

While working on Jupyter Notebook however, I did run into a few coding issues where I couldn't figure out how to do something in particular. An example of this was when I wanted to create a pivot table for my heatmap. Searching for the answer on the internet, I discovered Stack Overflow ("Stack Overflow - Where Developers Learn, Share, & Build Careers"), a great resource for troubleshooting coding issues. It helped me out with my project quite a lot.

Conclusion

After completing my project, I was quite happy with the analysis, as I did intuitively think that there would be linear increase in the rate of transformative drugs simply due to the rates of technological innovation exploding in recent years. It was good to see that intuition visualised. I do feel however, that this question can be further expanded upon as for possible future analysis, I would look into other contributing factors such as funding and the rate at which everyday drugs are approved.

Bibliography:

“FDA Breakthrough Drugs list 2012-Current.” *FDA,* https://www.fda.gov/media/95302/download

“Stack Overflow - Where Developers Learn, Share, & Build Careers.” *Stack Overflow*, https://stackoverflow.com/. Accessed 13th Apr. 2025.