SHREYA JOSHI

ALY 6980

Module 10 Assignment

Individual Project Draft Proposal

Prof. Roy Wada

INTRODUCTION

The primary aim of my individual project is to analyze the NIH spending data, specifically in the context of Dravet Syndrome. This project is an integral part of our larger group endeavor, which seeks to provide comprehensive insights into how federal research funding translates into tangible health benefits. By examining detailed expenditure data and correlating it with health outcomes, I can identify key factors that drive successful health interventions and policy implementations, ultimately aiding our group in delivering a robust final analysis to our sponsor.

Dravet Syndrome is a rare and severe type of epilepsy that typically starts in infancy. It is marked by prolonged seizures, developmental delays, and a significantly increased risk of sudden unexpected death in epilepsy (SUDEP). Understanding the allocation and efficacy of NIH funding in research related to this condition is critical. According to Smith (2017), traditional linear regression analysis often falls short in capturing the complex relationships present in social network data, which can also be analogous to the complex funding patterns and health outcomes seen in NIH-funded research. Therefore, I propose integrating linear regression with dimension reduction techniques and correlation analysis to unravel the multifaceted connections within our dataset. These methods will enable us to better understand how various financial inputs from the NIH influence diverse health outcomes.

Our sponsor, the GRIK THERAPEUTICS, dedicated to advancing medical research and improving public health especially in pediatric patients with epilepsy. The dataset provided by the NIH includes comprehensive information on funded projects, encompassing variables such as project titles, public health relevance, administering ICs (Institutes and Centers), fiscal years, and total costs. This rich dataset offers an excellent opportunity to perform a robust analysis that aligns with our project's objectives. By leveraging this data, I aim to draw meaningful correlations between NIH spending and the progression of research in Dravet Syndrome.

Previous research, such as the work by Johnson (2021), suggests that in the context of social network analysis of political alliances, dimension reduction is crucial due to the exponential growth of network nodes. Similarly, our dataset contains numerous financial and categorical variables, necessitating dimension reduction techniques to manage and interpret high-dimensional data effectively. This approach will help us identify significant patterns and trends in NIH funding allocations and their subsequent impacts on research outcomes.

Our approach begins with a detailed examination of Direct and Indirect Costs, which represent significant components of overall project expenditures. Direct Costs typically include expenses directly associated with the research, such as salaries, equipment, and supplies, while Indirect Costs cover overhead expenses like facility maintenance and administrative support. By quantifying the relationship between these costs and the Total Cost, I aim to identify key drivers of project expenses and provide insights into funding distribution patterns.

The importance of this analysis is further underscored by the potential real-world implications. By understanding how NIH funds are distributed and which factors contribute most significantly to successful research outcomes, I can provide actionable insights to NIH policymakers. These insights can help optimize funding strategies, ensuring that resources are allocated in ways that maximize public health benefits, particularly for conditions as severe and impactful as Dravet Syndrome.

In summary, the purpose of my individual project is to apply advanced statistical techniques to the NIH spending data to identify the key drivers of successful health outcomes. This proposal outlines a structured approach to our research, justifying each step with references from the literature. The ultimate goal is to deliver a comprehensive analysis that supports our group's overarching objective of enhancing the effectiveness of NIH funding in improving public health. Through meticulous analysis and integration of advanced methodologies, this project will contribute significantly to our final group deliverable, offering valuable insights for stakeholders involved in health research funding.

METHODS

Our research methodology integrates advanced statistical techniques with machine learning algorithms to analyze the NIH spending data and evaluate its impact on public health outcomes, specifically focusing on Dravet Syndrome. The primary methods employed include linear regression analysis, dimension reduction techniques, and Random Forest regression. These methods are chosen to address the complexity of the dataset and to uncover meaningful insights into NIH funding patterns and their correlation with research outcomes.

I will begin by conducting a linear regression analysis to understand the relationship between Direct Costs, Indirect Costs, and Total Cost of NIH-funded projects. This analysis aims to quantify the extent to which each type of cost influences the overall project expenditure. I will utilize software packages such as Python with libraries like NumPy, pandas, and scikit-learn to perform the regression analysis. The choice of linear regression is justified by its widespread use in analyzing the impact of independent variables on a dependent variable, as noted by Smith (2017) in the context of complex social network data analysis.

To further validate our regression analysis and evaluate the predictive power of various features, I will implement a Random Forest regression model. Random Forest is a powerful ensemble learning technique known for its ability to manage nonlinear interactions and relationships among variables effectively. This approach will help us assess the importance of different features in predicting Total Cost and provide insights into the key drivers of NIH funding allocation. The Random Forest algorithm will be implemented using Python, specifically leveraging libraries such as scikit-learn for model building and evaluation.

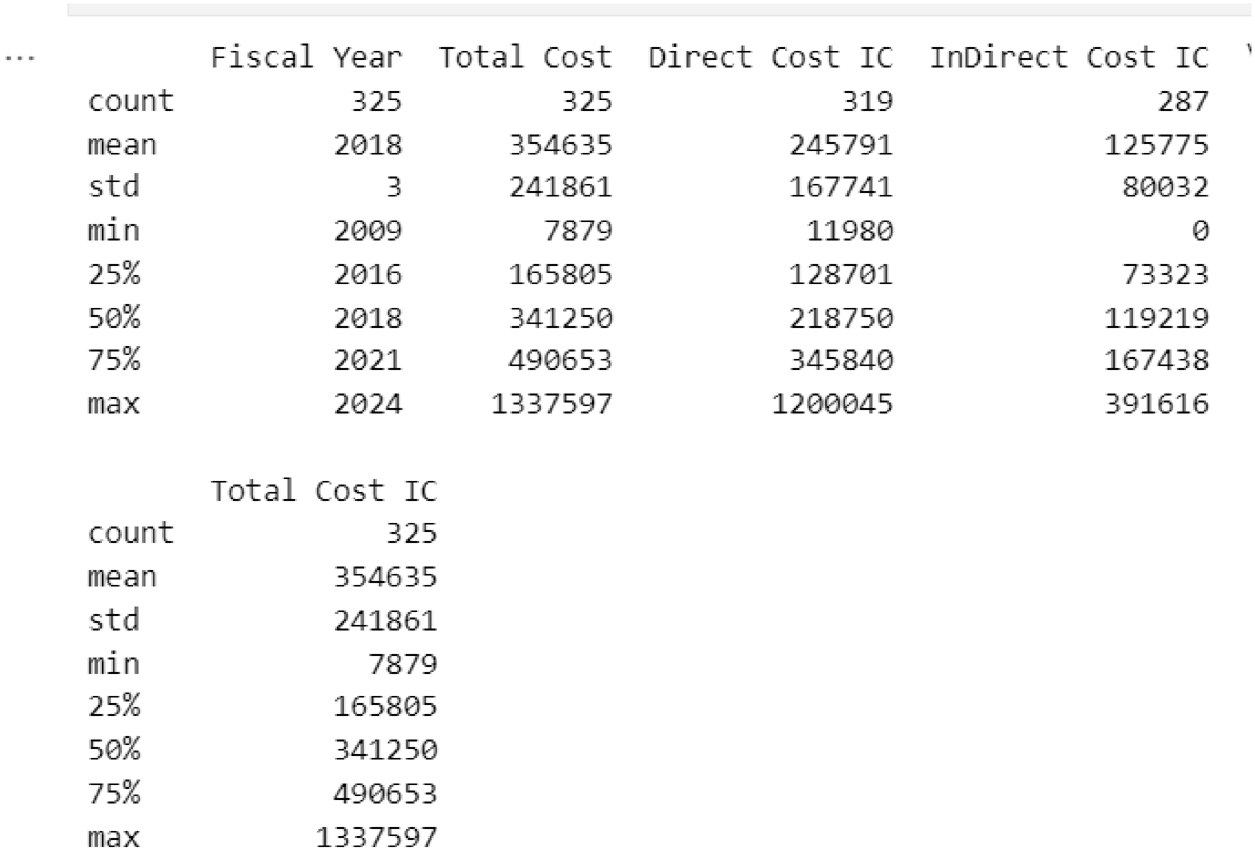
For data preprocessing, analysis, and modeling, I will primarily use Python programming language along with relevant libraries and frameworks. The commands and libraries mentioned above will be utilized to execute the respective methodologies, ensuring a systematic and rigorous approach to our research. I also have created one PowerBI dashboard to visualize Funds and investors data for pediatric epilepsy pharma research companies from Pitchbook Data.

PRELIMINARY RESULTS

In this section, I provide an in-depth analysis of the NIH spending data, showcasing summary statistics, visualizations, correlation analysis, and regression results. These preliminary results lay the groundwork for further analysis and insights in the subsequent stages of the project.

Summary Statistics

Summary statistics provide a comprehensive overview of the key variables in the dataset, helping us understand the central tendencies, dispersion, and range of the data. Here, I focus on the Direct Cost IC, InDirect Cost IC, and Total Cost variables.



```
...        Fiscal Year  Total Cost  Direct Cost IC  InDirect Cost IC  \
    count          325         325             319               287
    mean          2018      354635          245791            125775
    std              3      241861          167741             80032
    min           2009        7879           11980                 0
    25%           2016      165805          128701             73323
    50%           2018      341250          218750            119219
    75%           2021      490653          345840            167438
    max           2024     1337597         1200045            391616

           Total Cost IC
    count            325
    mean          354635
    std           241861
    min             7879
    25%           165805
    50%           341250
    75%           490653
    max          1337597
```

The summary statistics provide a comprehensive overview of key variables in the NIH spending dataset, focusing on Direct Cost IC, InDirect Cost IC, and Total Cost. The count indicates the number of non-missing observations for each variable, ensuring the robustness of our analysis. The mean represents the average value of the costs, giving us a central point around which the data is distributed. The standard deviation measures the dispersion or variability of the costs, highlighting how spread out the values are from the mean. The minimum and maximum values show the range of the data, with the lowest and highest observed values respectively. The 25th percentile indicates the value below which 25% of the observations fall, providing insight into the lower quartile of the data. The median, or 50th percentile, splits the dataset into two equal halves, offering a central value that is less affected by outliers compared to the mean. Finally, the

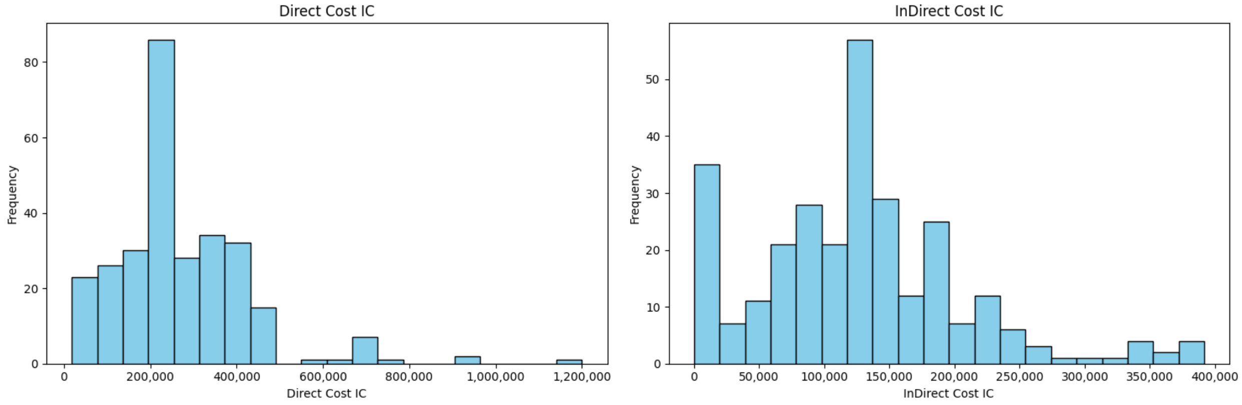
75th percentile marks the value below which 75% of the observations fall, helping us understand the upper quartile distribution of the costs. These statistics reveal significant variability in NIH funding, with some projects receiving over a million dollars while others receive much less, indicating a right-skewed distribution with a few high-value outliers.

Visualizations

Visualizations help us to better understand the distribution and relationships within the data. Here, I present histograms and box plots for Direct Cost IC and InDirect Cost IC.

Histogram of Direct and Indirect Costs

Histograms show the frequency distribution of Direct and Indirect Costs, allowing us to observe patterns such as skewness and the presence of outliers.

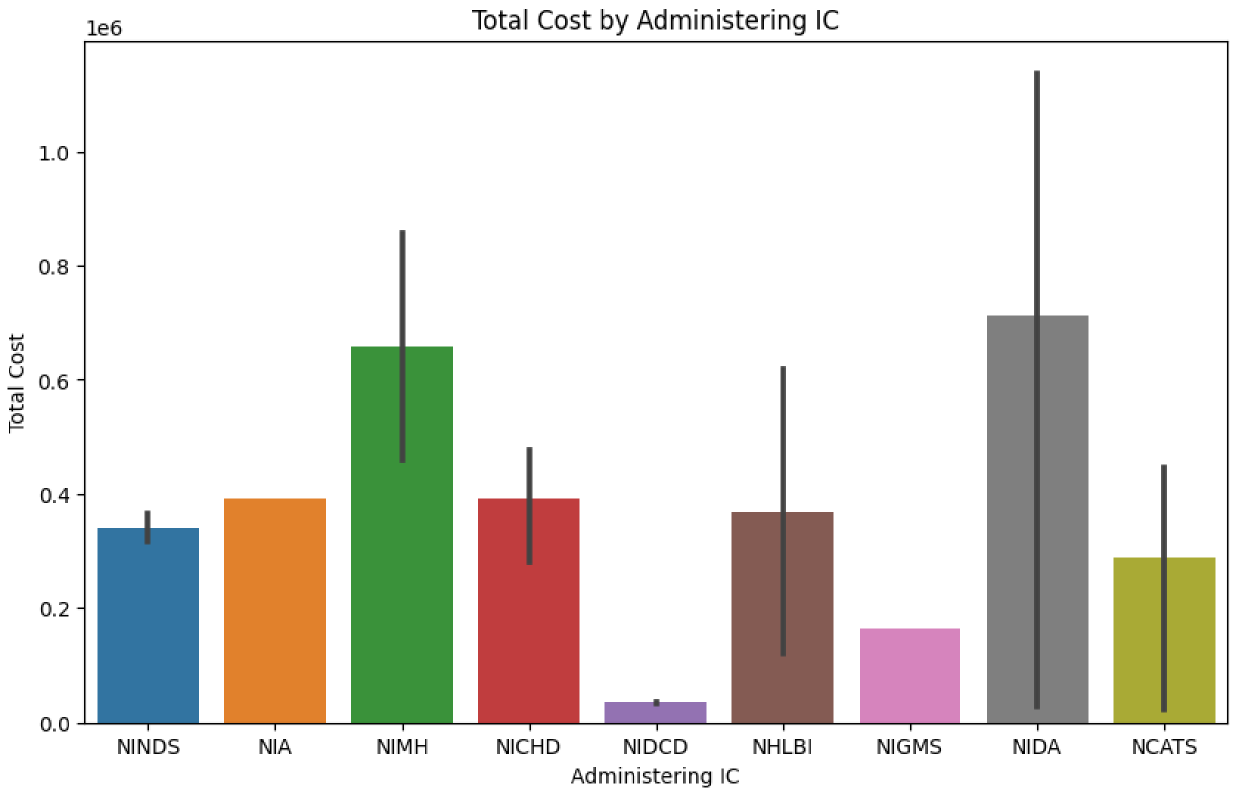


The histograms indicate that both Direct and Indirect Costs are right-skewed, with most projects receiving lower amounts of funding and a few projects receiving very high amounts.

Total Cost Analysis Across NIH Administering Integrated Centers

The bar chart displays the Total Cost across different Administering Integrated Centers (ICs) at the National Institutes of Health (NIH). The National Institute on Drug Abuse (NIDA) has the highest Total Cost, represented by the tallest bar, indicating a substantial allocation of funds to research and programs related to drug abuse. The National Institute of General Medical Sciences (NIGMS) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDCD) also have relatively high Total Costs, suggesting significant investments in their respective areas of focus.

Moderate Total Costs are observed for the National Institute of Mental Health (NIMH), National Institute on Aging (NIA), and National Institute of Neurological Disorders and Stroke (NINDS). The National Institute of Child Health and Human Development (NICHD) and the National Center for Advancing Translational Sciences (NCATS) have comparatively lower Total Costs. Notably, the National Heart, Lung, and Blood Institute (NHLBI) has the lowest Total Cost among the ICs depicted in the chart, implying a potentially smaller budget allocation or a narrower scope of research activities.



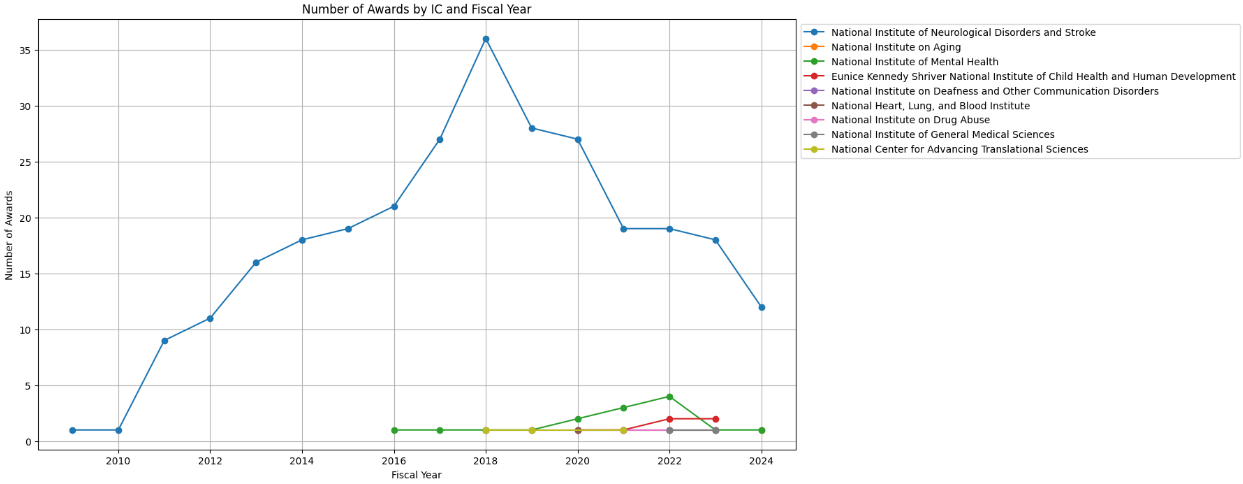
Number of Awards by IC and Fiscal Year

The graph illustrates the distribution of awards from various National Institutes of Health (NIH) institutes and centers over fiscal years 2010 to 2024. The number of awards is plotted on the y-axis, with fiscal years on the x-axis. The National Institute of Neurological Disorders and Stroke (NINDS) consistently leads with the highest award count, peaking notably around 2018-2019.

In contrast, the National Institute on Aging (NIA) and the National Institute of Mental Health (NIMH) show a lower number of awards, though they remain steady contributors over the years. Institutes such as the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the National Institute on Deafness and Other Communication Disorders (NIDCD), the National Heart, Lung, and Blood Institute (NHLBI), the National Institute on Drug Abuse (NIDA), the National Institute of General Medical Sciences (NIGMS), and the National

Center for Advancing Translational Sciences (NCATS) have fewer awards compared to NINDS, NIA, and NIMH.

The graph reveals a general trend of increasing awards until around fiscal years 2018-2019, followed by a decline in subsequent years for most institutes and centers, particularly NINDS. It is crucial to note that this graph provides a high-level overview, and additional context and information would be necessary to interpret the specific reasons behind the trends and variations observed across different institutes and centers at the NIH.



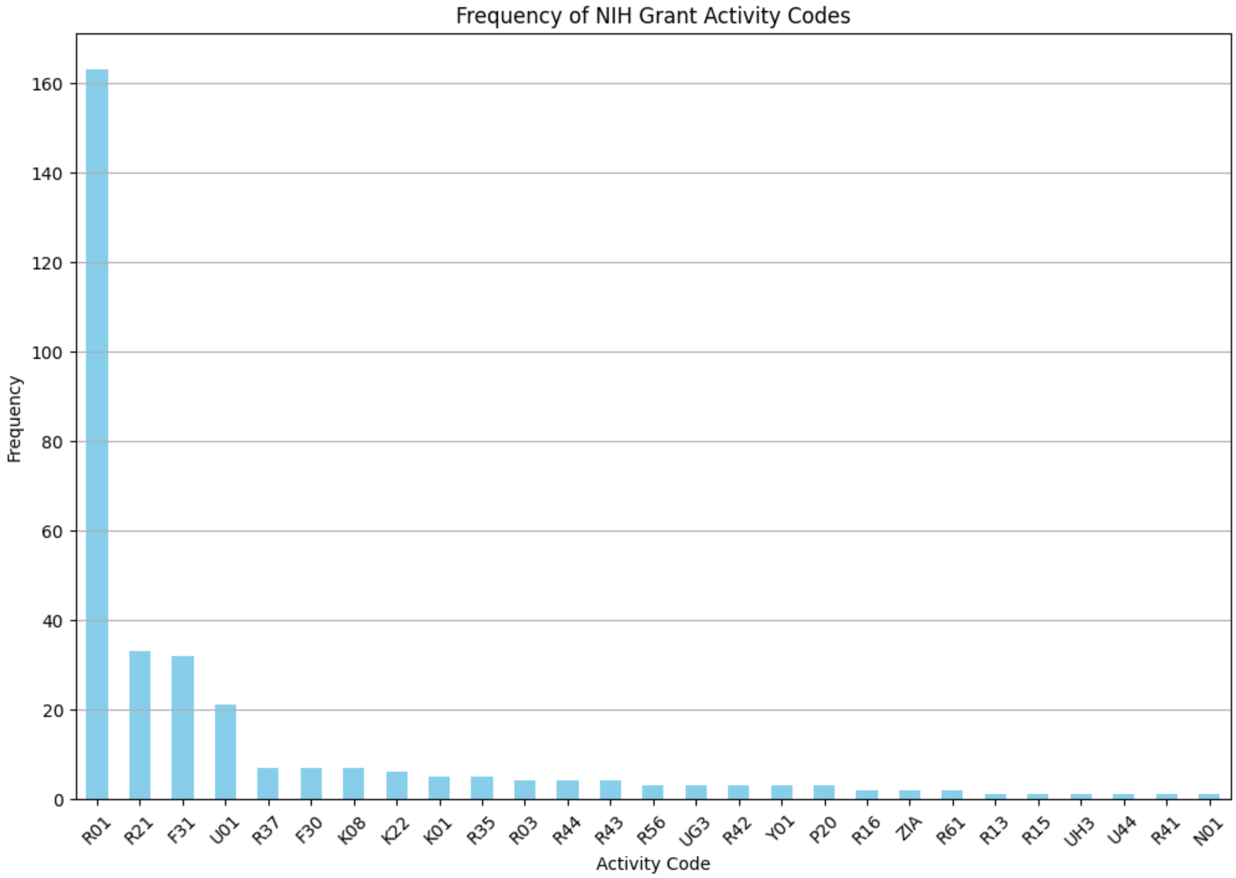
Interpretation of NIH Grant Activity Codes Frequency

The bar chart provides an insightful overview of the frequency distribution of NIH Grant Activity Codes, highlighting the prominence and utilization of various grant types within the NIH funding ecosystem. The R01: Research Project Grant stands out as the most frequently awarded grant, with a total of 160 occurrences, signifying its critical role in supporting research projects across a wide array of scientific disciplines. Following the R01, the R21: Exploratory/Developmental Grant and F31: Predoctoral Individual National Research Service Award also show substantial frequencies of around 40 and 35, respectively, indicating their importance in fostering exploratory research and supporting predoctoral training.

The U01: Research Project Cooperative Agreement is also notable with a frequency of approximately 30, reflecting its significance in facilitating collaborative research initiatives. Other grant types such as R37, F30, K08, K22, K01, and R35 exhibit moderate frequencies ranging from 10 to 20, suggesting their targeted applications for specific research and career development purposes.

In contrast, several grant types like the R03: Small Grant Program, R44: Small Business Innovation Research (SBIR) Phase II, and R43: Small Business Innovation Research (SBIR) Phase I are less common, with frequencies under 10, indicating their specialized nature or limited scope. Numerous activity codes, including R16, ZIA, R61, R13, R15, among others, display very low frequencies (5 or less), underscoring their specialized or niche roles within the NIH funding structure.

Overall, the distribution is highly skewed, with a few grant types, particularly the R01, dominating the NIH funding landscape. This pattern underscores the prioritization and widespread application of certain grant mechanisms over others, reflecting the strategic funding decisions and research priorities within the NIH.



Number of Awards per IC and Fiscal Year

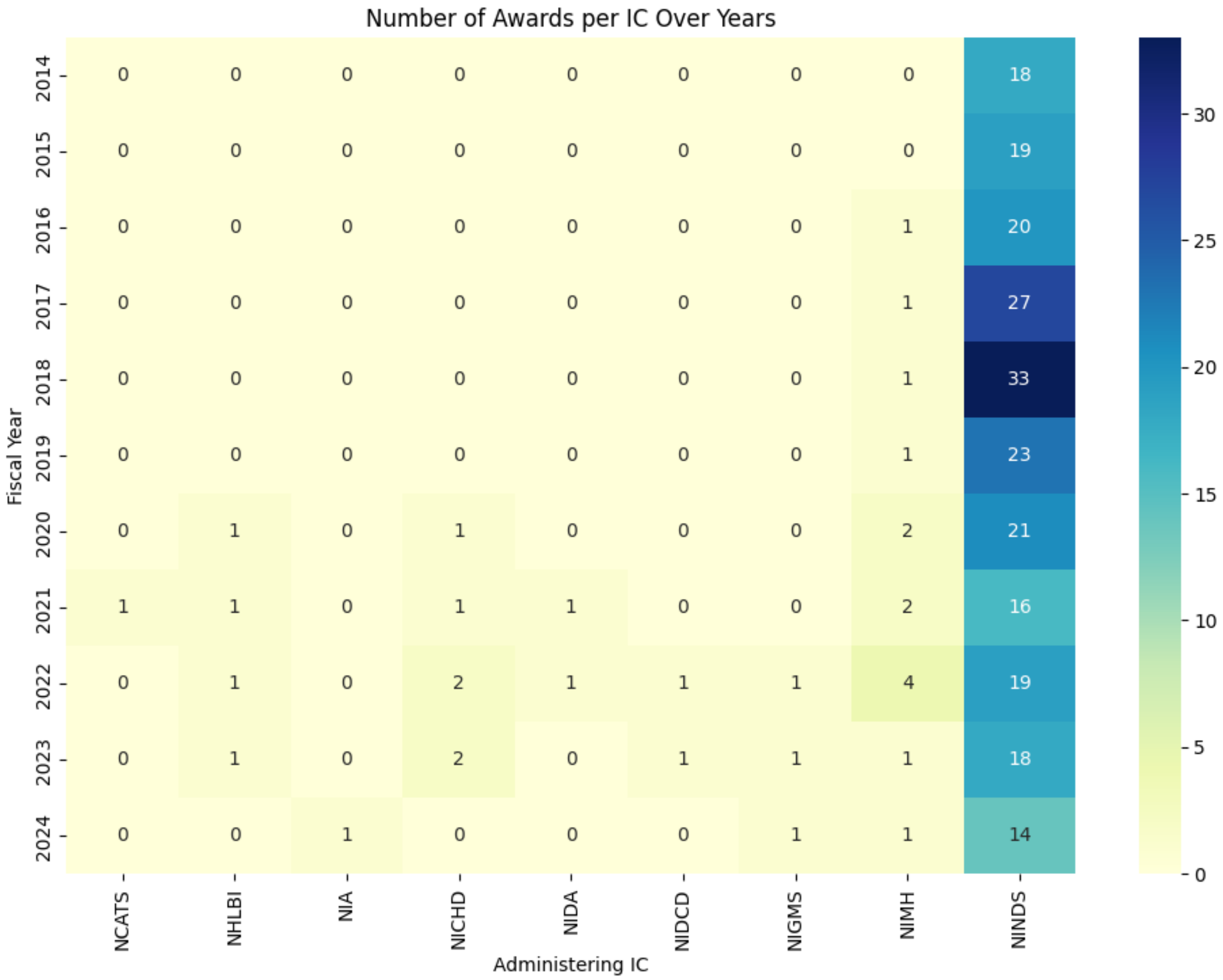
The heatmap illustrates the distribution of awards from various National Institutes of Health (NIH) institutes and centers over fiscal years 2014 to 2024. The number of awards is represented by color intensity, with lighter shades indicating fewer awards and darker shades indicating more awards.

From 2014 to 2019, most institutes show no awards, reflecting either inactivity or data absence. Starting in 2020, awards begin to appear across several institutes, showing a more active funding distribution. Notably, the National Institute of Neurological Disorders and Stroke (NINDS) shows the highest activity, especially with a peak in awards around 2018-2019. The National

Institute of Mental Health (NIMH) and the National Institute on Aging (NIA) also show steady contributions, though with lower numbers compared to NINDS.

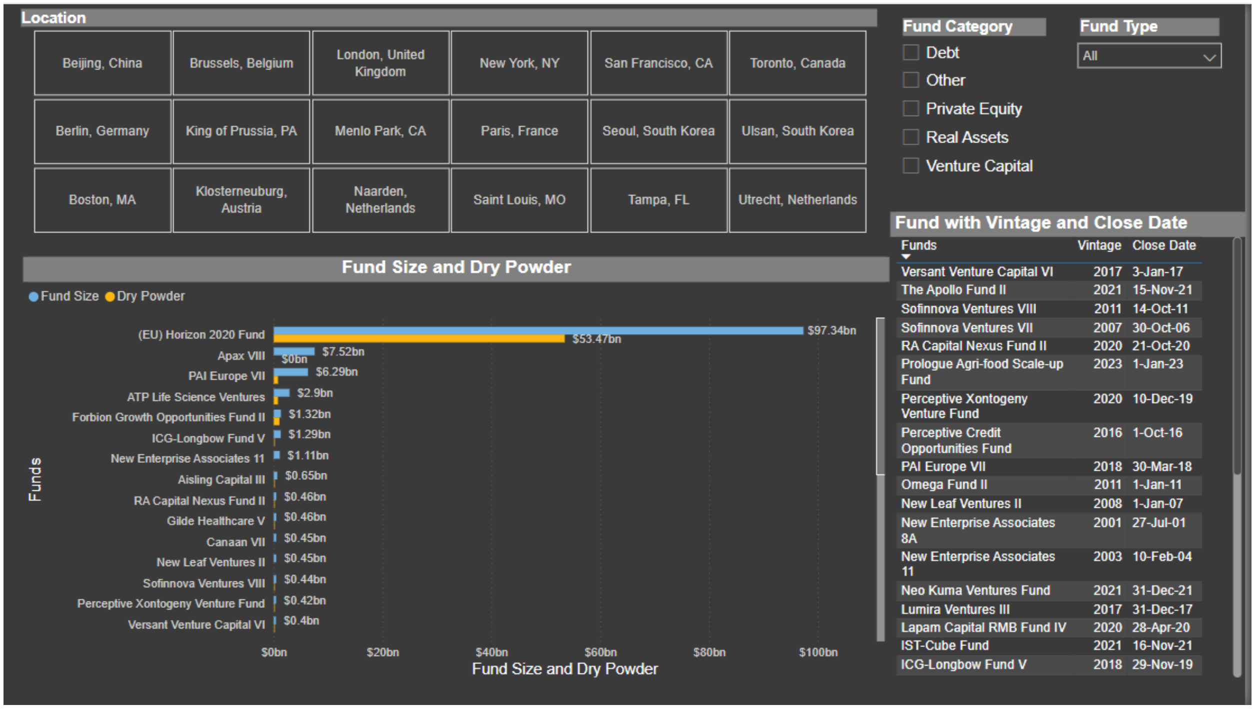
In the more recent years, from 2020 to 2024, an increase in awards is seen for multiple institutes such as NIA, NIMH, and NINDS. There is noticeable activity in 2023 and 2024, with some institutes like the National Heart, Lung, and Blood Institute (NHLBI), the National Institute on Drug Abuse (NIDA), and the National Institute on Deafness and Other Communication Disorders (NIDCD) also showing awards. The color scale on the right provides a reference for the number of awards, with the darkest color representing the highest number of awards (33).

This heatmap reveals trends and shifts in the distribution of NIH awards over time, highlighting key periods of increased funding activity and the prominence of certain institutes in receiving awards. Additional context and information would be necessary to interpret the specific reasons behind the trends and variations observed across different institutes and centers at the NIH.



Dynamic PowerBI Interactive Dashboard:

The dashboard serves as a pivotal tool for our sponsor, GRIK THERAPEUTICS, providing a comprehensive and dynamic visualization of Pitchbook funding data. By offering the capability to adjust parameters such as location, fund type, fund category, vintage year, and latest close date, the dashboard enables users to gain nuanced insights into funding patterns. This flexibility is crucial for stakeholders seeking to make informed decisions based on the availability and distribution of funds in the market.



Location

Beijing, China	Brussels, Belgium	London, United Kingdom	New York, NY	San Francisco, CA	Toronto, Canada
Berlin, Germany	King of Prussia, PA	Menlo Park, CA	Paris, France	Seoul, South Korea	Ulsan, South Korea
Boston, MA	Klosterneuburg, Austria	Naarden, Netherlands	Saint Louis, MO	Tampa, FL	Utrecht, Netherlands

Fund Category

□ Debt
□ Other
□ Private Equity
□ Real Assets
□ Venture Capital

Fund Type: All

Fund Size and Dry Powder

Fund with Vintage and Close Date

Funds	Vintage	Close Date
Versant Venture Capital VI	2017	3-Jan-17
The Apollo Fund II	2021	15-Nov-21
Sofinnova Ventures VIII	2011	14-Oct-11
Sofinnova Ventures VII	2007	30-Oct-06
RA Capital Nexus Fund II	2020	21-Oct-20
Prologue Agri-food Scale-up Fund	2023	1-Jan-23
Perceptive Xontogeny Venture Fund	2020	10-Dec-19
Perceptive Credit Opportunities Fund	2016	1-Oct-16
PAI Europe VII	2018	30-Mar-18
Omega Fund II	2011	1-Jan-11
New Leaf Ventures II	2008	1-Jan-07
New Enterprise Associates 8A	2001	27-Jul-01
New Enterprise Associates 11	2003	10-Feb-04
Neo Kuma Ventures Fund	2021	31-Dec-21
Lumira Ventures III	2017	31-Dec-17
Lapam Capital RMB Fund IV	2020	28-Apr-20
IST-Cube Fund	2021	16-Nov-21
ICG-Longbow Fund V	2018	29-Nov-19

Predictive Modeling

```
Linear Regression Model:
Mean Squared Error: 17984246513607.816
R-squared: -394.7024745610306

Random Forest Model:
Mean Squared Error: 15543410331.302656
R-squared: 0.6580025787257692

XGBoost Model:
Mean Squared Error: 8403432984.17562
R-squared: 0.8151015543447976
```

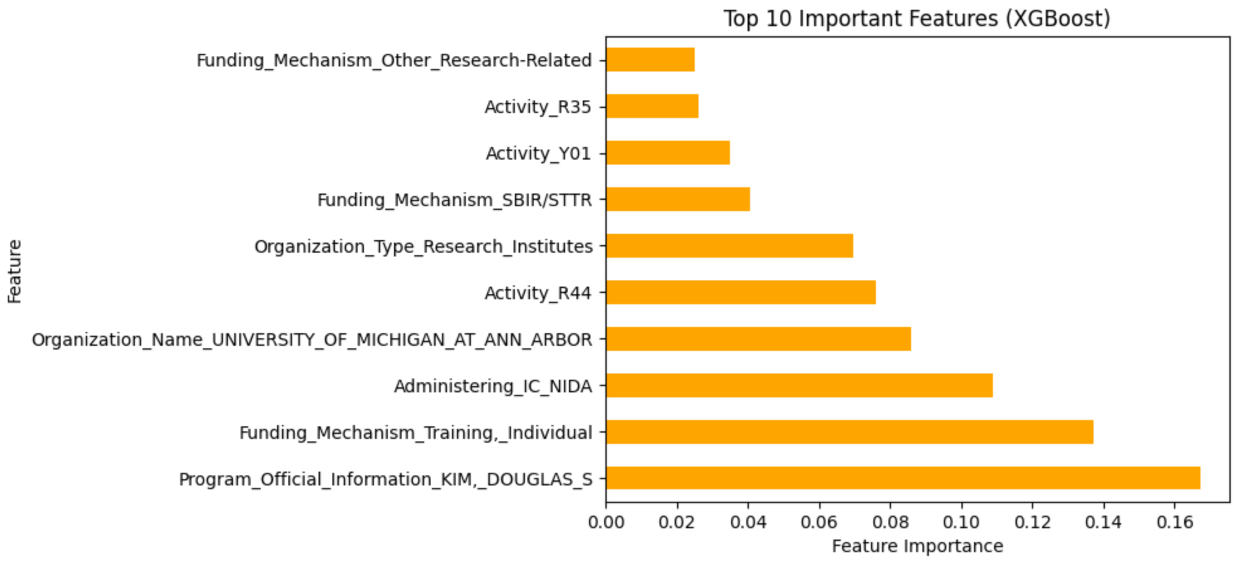
The data analysis reveals several key insights regarding the performance of different predictive models for NIH funding. Initially, the linear regression model performed poorly, evidenced by a mean squared error (MSE) of approximately 17.98 trillion and an R-squared value of -394.7. This negative R-squared indicates the model is significantly worse than the mean prediction.

Conversely, the random forest model performed substantially better, with an MSE of 15.54 billion and an R-squared value of 0.658, suggesting a reasonably good fit.

Upon performing cross-validation, the random forest model demonstrated better performance, with an MSE of 18.46 billion and an R-squared of 0.683, reinforcing its robustness compared to the linear regression model. Feature importance analysis using the random forest model highlighted significant predictors of funding. The top predictors included specific funding mechanisms, organizational types, project activities, and geographical attributes. For example, features such as "Funding Mechanism_Training, Individual," "Organization Type_Research Institutes," and specific project activities like "Activity_R01" and "Activity_R35" significantly influenced funding outcomes. These insights suggest that projects categorized under training mechanisms typically receive less funding compared to research-focused grants. Additionally, established research institutions and particular locations showed positive impacts on funding, emphasizing the importance of organizational reputation and geographic advantages in securing NIH funding.

The XGBoost model outperformed the random forest model, with an MSE of 8.4 billion and an R-squared value of 0.815, indicating a very strong fit. This model's performance shows that it explains around 81.5% of the variance in funding amounts, making it a highly reliable predictor. The strong performance of the XGBoost model underscores its utility in capturing complex relationships in the data, providing valuable insights for decision-making in the allocation of NIH funding.

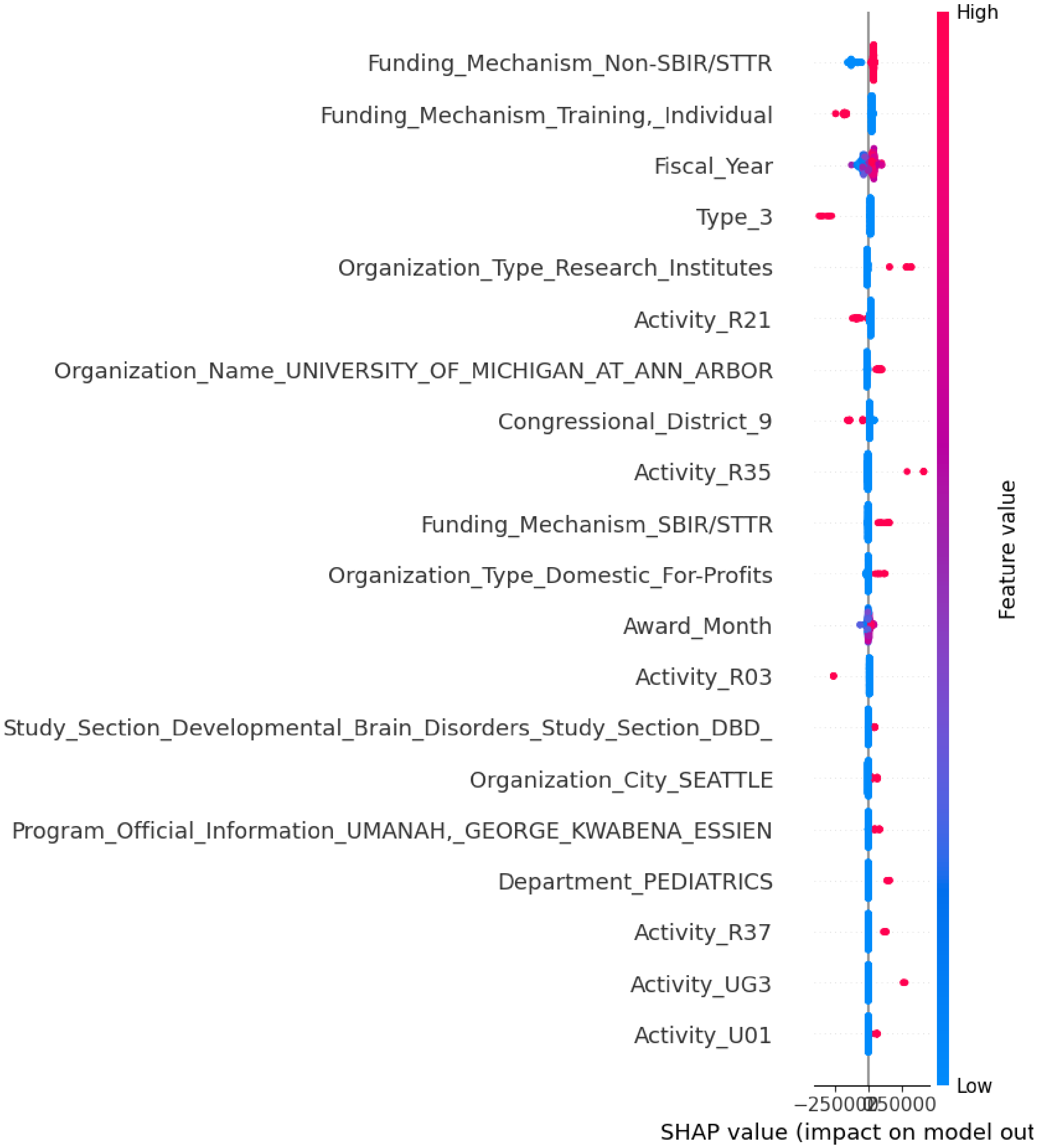
In conclusion, the comparative analysis of these models highlights the superior predictive power of the XGBoost model for NIH funding data, followed by the random forest model. The poor performance of the linear regression model emphasizes the need for more sophisticated approaches to capture the underlying patterns in the data. These findings can guide strategic decisions for research institutions and grant applicants, helping them to better tailor their proposals and improve their chances of securing funding.



The feature importance plot from the XGBoost model provides a clear visualization of the top factors influencing NIH funding allocations, highlighting key aspects for GRIK THERAPEUTICS to consider in their strategic planning. The plot ranks features based on their importance to the model's predictive power. Notably, the feature "Program_Official_Information_KIM_DOUGLAS_S" has the highest importance, suggesting that the involvement of specific program officials can significantly affect funding outcomes. Other important features include "Funding_Mechanism_Training,_Individual" and

"Administering_IC_NIDA," indicating that individual training funding mechanisms and administration by the National Institute on Drug Abuse are crucial factors.

Additionally, the University of Michigan at Ann Arbor appears prominently, reflecting the institution's strong influence on funding success. Research institutes and specific activities like R44 and R35 also play vital roles. Understanding these influential features allows GRIK THERAPEUTICS to tailor their grant applications more effectively, emphasizing areas that align with the key drivers identified by the model. By focusing on these important aspects, the organization can enhance their funding strategies, improve their chances of securing NIH grants, and advance their research goals efficiently.



The SHAP (SHapley Additive exPlanations) value plot displayed in the dashboard provides a detailed interpretation of the key factors influencing NIH funding allocations from a business perspective. The plot highlights the most influential features affecting the funding outcomes, such as funding mechanisms, organization types, fiscal year, and specific research activities. For instance, non-SBIR/STTR and individual training funding mechanisms, as well as research institutes and specific activities like R21 and R35, have significant impacts on the funding amounts. The feature values are color-coded, with higher values in pink and lower values in blue, showing the positive or negative contributions to the model's predictions. This detailed analysis helps GRIK THERAPEUTICS identify which factors are most critical for securing NIH funding, allowing the organization to strategically focus on strengthening these aspects in their grant applications. Understanding these key drivers can enhance their funding strategies, ensuring that resources are directed towards the most impactful areas, ultimately improving their chances of successful funding and advancing their research initiatives.

Business Implementation

1.Strategic Grant Application Focus: Based on the analysis, prioritize grant applications that align with highly influential factors identified by the XGBoost model, such as specific funding mechanisms (e.g., R01 grants), organizational types (e.g., research institutes), and activities (e.g., R35 projects). This strategic alignment will enhance the competitiveness of grant proposals.

2.Enhanced Collaboration Strategy: Leverage insights from the heatmap and award distribution graphs to foster collaborations with institutes like the National Institute of Neurological Disorders and Stroke (NINDS), which consistently receive high award counts. This approach can facilitate joint research efforts and increase access to NIH funding.

3.Geographical Strategy: Focus on locations identified as influential in securing NIH funding, such as the University of Michigan at Ann Arbor. Strengthening partnerships or establishing satellite operations in these regions could optimize funding opportunities and support research initiatives effectively.

4.Resource Allocation Optimization: Utilize the PowerBI dashboard to dynamically visualize and analyze Pitchbook funding data. Adjust parameters like fund type and vintage year to strategically allocate resources and investments in pediatric epilepsy research, ensuring alignment with current market trends and funding availability.

Conclusion

The analysis of NIH spending data using advanced statistical techniques has provided actionable insights specific to optimizing funding strategies for Dravet Syndrome research. By leveraging the predictive power of the XGBoost model and strategic insights from the PowerBI dashboard, GRIK THERAPEUTICS can focus on securing grants that align with influential factors such as specific funding mechanisms (e.g., R01 grants), organizational types (e.g., research institutes), and key geographical locations (e.g., University of Michigan at Ann Arbor). These findings enable GRIK THERAPEUTICS to refine its grant application approach, enhance collaboration efforts with high-activity institutes like NINDS, and strategically allocate resources for maximum impact in advancing pediatric epilepsy treatments.

REFERENCES

1. Brunklaus A, Pérez-Palma E, Ghanty I, Xinge J, Brilstra E, Ceulemans B, Chemaly N, de Lange I, Depienne C, Guerrini R, Mei D, Møller RS, Nabbout R, Regan BM, Schneider AL, Scheffer IE, Schoonjans AS, Symonds JD, Weckhuysen S, Kattan MW, Zuberi SM, Lal D. Development and Validation of a Prediction Model for Early Diagnosis of SCN1A-Related Epilepsies. Neurology. 2022 Mar 15;98(11):e1163-e1174. doi: 10.1212/WNL.0000000000200028. Epub 2022 Jan 24. PMID: 35074891; PMCID: PMC8935441.
2. Feng T, Makiello P, Dunwoody B, Steckler F, Symonds JD, Zuberi SM, Dorris L, Brunklaus A. Long-term predictors of developmental outcome and disease burden in SCN1A-positive Dravet syndrome. Brain Commun. 2024 Jan 9;6(1):fcae004. doi: 10.1093/braincomms/fcae004. PMID: 38229878; PMCID: PMC10789590.
3. Wallace ML, Mentch L, Wheeler BJ, Tapia AL, Richards M, Zhou S, Yi L, Redline S, Buysse DJ. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med Res Methodol. 2023 Jun 19;23(1):144. doi: 10.1186/s12874-023-01965-x. PMID: 37337173; PMCID: PMC10280951.
4. Wardrope A, Jamnadas-Khoda J, Broadhurst M, Grünewald RA, Heaton TJ, Howell SJ, Koepp M, Parry SW, Sisodiya S, Walker MC, Reuber M. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. Neurol Clin Pract. 2020 Apr;10(2):96-105. doi: 10.1212/CPJ.0000000000000726. PMID: 32309027; PMCID: PMC7156196.

5. National Institutes of Health. (n.d.). Activity Codes Search Results. Retrieved from https://grants.nih.gov/grants/funding/ac_search_results.htm

6. Sayfiddinov, D. (2018, June 6). How to build a dynamic Power BI reporting dashboard. Practical 365. Retrieved from https://practical365.com/dynamic-power-bi-reporting-dashboard/