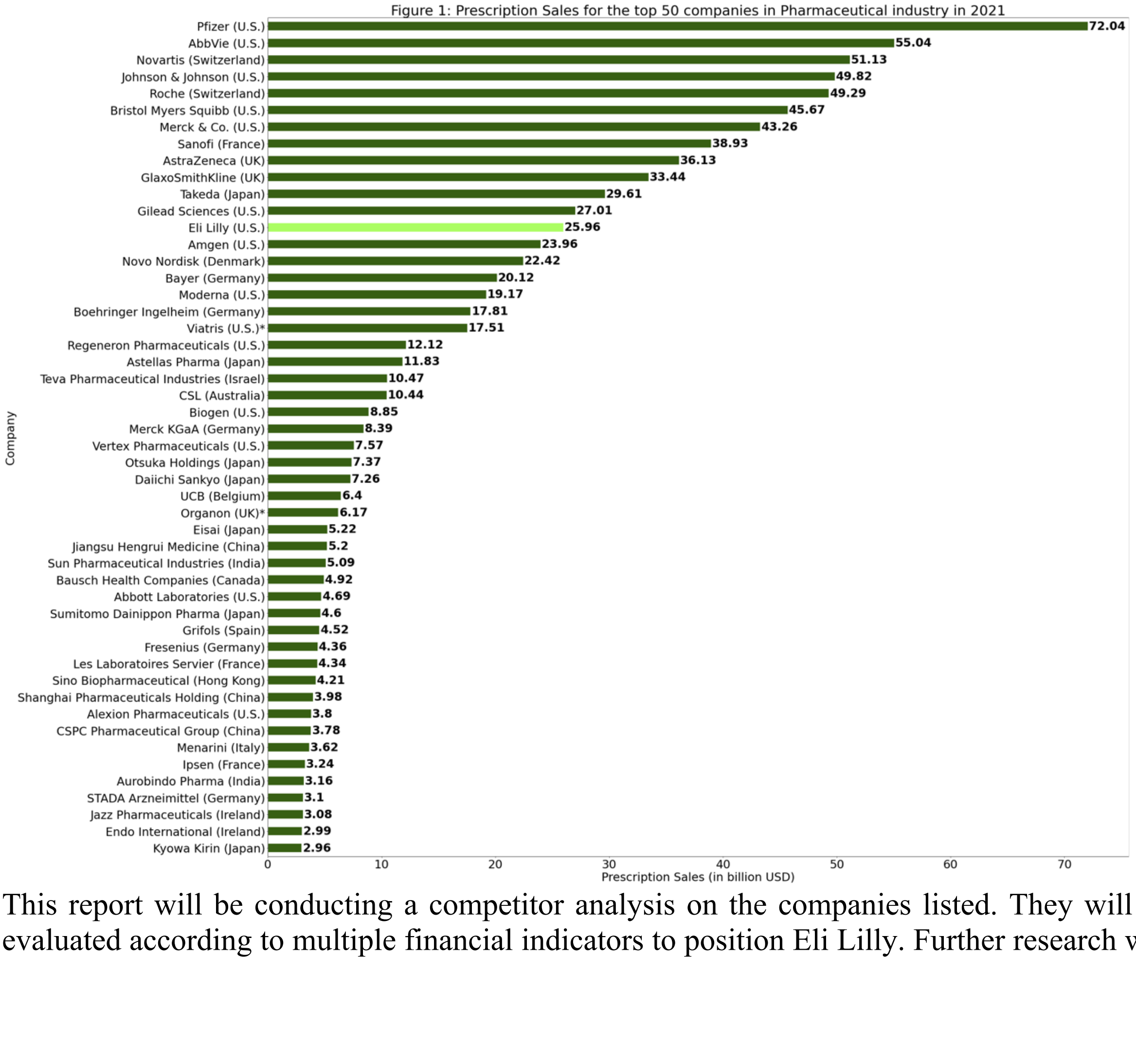
# R&D and Pharma: Does Spending on Research Boost Profit for the Pharmaceutical Industry?

## Introduction

Scherer states that the pharmaceutical industry is "one of the world's most research-intensive industries, generating a continuous stream of new products that save lives and raise the quality of life" (Scherer, 2000). Manufacturers tie up with researchers in universities and national laboratories to develop new pharmaceutical offerings. These new products are then protected using patents and then sold to recoup the money sunk into research and generate profit for the manufacturers. The process of research, development and clinical testing is arduously long and money intensive. The industry is subject to more government regulation and public policy than most other sectors (Lakdawalla, 2018). Pharmaceuticals is an umbrella term referring to medicines, vaccines, and other drugs (Scherer, 2000). Due to the sensitive nature of the products, these are mostly regulated and purchased when prescribed by a physician and are strictly regulated.

One of the measures of financial performance is the sale of the products. Figure 1 displays the prescription sales of the top pharmaceutical companies globally in 2021. According to Statista, Eli Lilly comes in 13th among pharmaceutical companies in 2021 with prescription sales valuing 25.96 billion USD. Eli Lilly also comes in 4th among pharmaceutical companies with a value of 292.8 billion USD by enterprise value (Executive, 2022). Eli Lilly was founded in 1876 and has been in the field of pharmaceuticals for over 140 years (Company Profiles: Eli Lilly & Company).



This report will be conducting a competitor analysis on the companies listed. They will be evaluated according to multiple financial indicators to position Eli Lilly. Further research will be done to evaluate expenditure on research and marketing to check whether fund allocation towards research is yielding results and whether the priority needs to shift towards marketing current product lines instead.

## Methods

*Data Collection*

The following sources of data were used to compile this report –

1. In order to evaluate the competitors of Eli Lilly, income statements of the competitors listed in Figure 1 were used (S&P Capital IQ (Firm), 2013). GlaxoSmithKline, Sumitomo Dainippon Pharma (Japan), Les Laboratoires Servier, Endo, Menarini and Shanghai Pharmaceuticals Holding (China) don't report R&D expenses so they were excluded.
2. Patent information was obtained from the Purple and Orange Book databases that US FDA uploads. The Purple Book contains "information about all FDA-licensed biological products regulated by CDER, including licensed biosimilar and interchangeable products, and their reference products, and FDA-licensed allergenic, cellular and gene therapy, hematologic, and vaccine products regulated by CBER" (Administration, 2022). The Orange Book "identifies drug products approved on the basis of safety and effectiveness by the Food and Drug Administration (FDA) under the Federal Food, Drug, and Cosmetic Act (the Act) and related patent and exclusivity information" ('Orange Book Data Files,' 2022).

The following variables were selected for further usage –

1. From income statement
    - 1.1. Year – year of income statement reporting
    - 1.2. Company – name of company
    - 1.3. Gross Profit – gross profit is the profit a company makes after deducting the cost of goods sold (COGS) from its revenue. It measures the efficiency of a company's production and pricing strategies (Kimmel, Weygandt and Kieso, 2007).
    - 1.4. Net Income – company's total profit after deducting all expenses, including taxes and interest, from its revenue. It is a measure of a company's overall profitability.
    - 1.5. Total Revenue – a measure of the assets received (for example, cash) in exchange for products sold by the company (Davidson *et al.*, 1979).
    - 1.6. Research And Development Expenses – the costs a company incurs in the process of developing new products or improving existing ones. Henceforth known as R&D expenses for the rest of the report.
    - 1.7. Selling and Marketing Expenses – the costs a company incurs to sell its products or services, including salaries and commissions for sales staff, marketing campaigns, and other expenses related to sales and marketing efforts.
2. From patent information
    - 2.1. Applicant Name – name of company applying for patent
    - 2.2. Patent Issue Date – date of issue of patent

*Data Cleaning*

Cleaning was done across three main aspects –

1) Statista data for top 50 companies by prescription sales –
    a) Columns were renamed appropriately.
2) Income Statements –
    a) These income statements were put through a pipeline to prepare the data
        i) Year column was renamed from "Recommended: S&P Capital IQ – Standard".
        ii) Dataset was transposed using Year as axis.
        iii) "FY" was removed from the years.
        iv) White spaces were removed from column names.
        v) All columns except for gross profit, net income, total revenue, R&D expenses, advertising expenses and selling and marketing expenses were dropped.
    b) Function was created to apply data cleaning pipeline to a dictionary with all the company names
    c) This was combined to form a single file containing the income statements for all companies across all the years that income statement was reported. There are 1183 rows present within this data frame.
3) Patent data –
    a) Data was imported from US FDA website for Orange Book and Purple Book databases.
    b) Only columns with approval date and applicant name were kept.
    c) Null values were filtered out.
    d) Date in rows with Year value as "Approved Prior to Jan 1, 1982" were changed to Dec 31, 1981.
    e) These were then aggregated by applicant name and year to get yearly data on applicants for each company.
4) Finally, a combined data frame was created with both income statements as well as patent data.

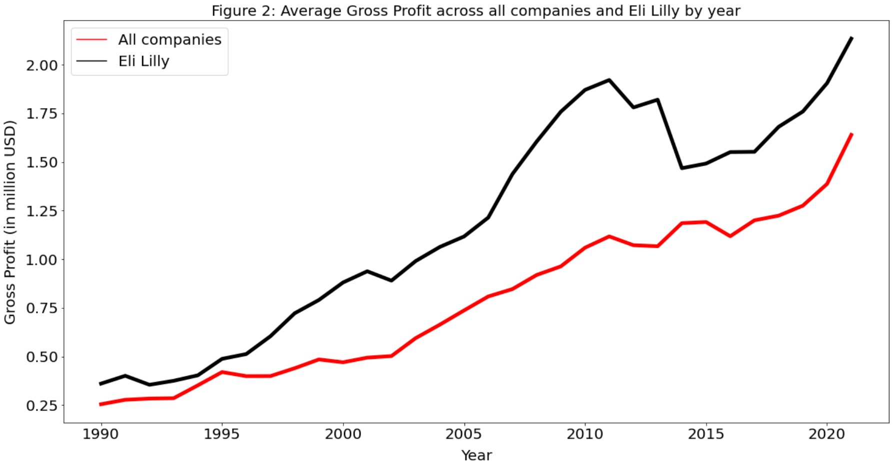
## *Data Description and Summary*

### *Gross Profit*

Based on Figure 2, it appears that Eli Lilly has consistently had higher gross profit compared to the average of all companies. In most years, Eli Lilly's gross profit was significantly higher than the average of all companies, indicating that the company has been more profitable in terms of gross profit compared to the average of all companies.

There are several potential reasons why Eli Lilly may have had higher gross profit compared to the average of all companies. These could include:

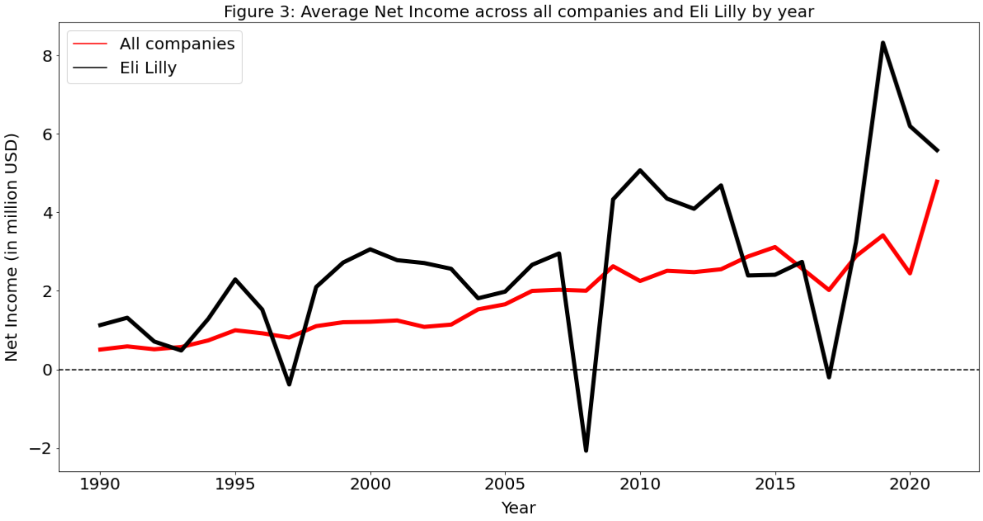
1. Higher prices for Eli Lilly's products: If Eli Lilly can charge higher prices for its products compared to its competitors, this could contribute to higher gross profit.
2. Greater efficiency in production: If Eli Lilly can produce its products more efficiently, it may be able to keep its COGS lower, leading to higher gross profit.



### *Net Income*

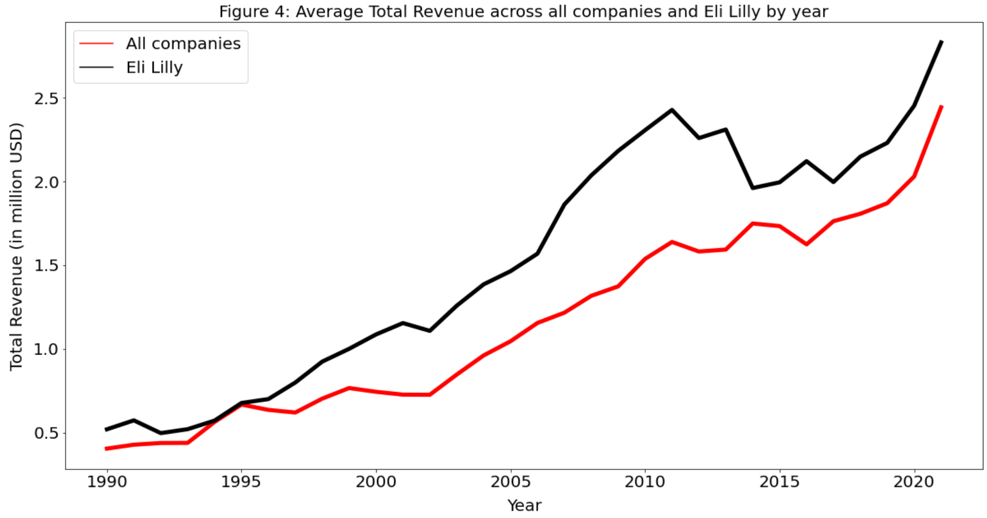
From Figure 3, Eli Lilly's net income has generally been higher than the average net income of all companies in the data. There are a few exceptions to this trend, such as in 1997, 2008 and 2017 when Eli Lilly's net income was significantly lower than the average.

A negative net income indicates that the company had more expenses than revenue, resulting in a loss for that year. Factors that can contribute to a negative net income include increased competition, changes in market conditions, and increased expenses for things like R&D, marketing, and production.



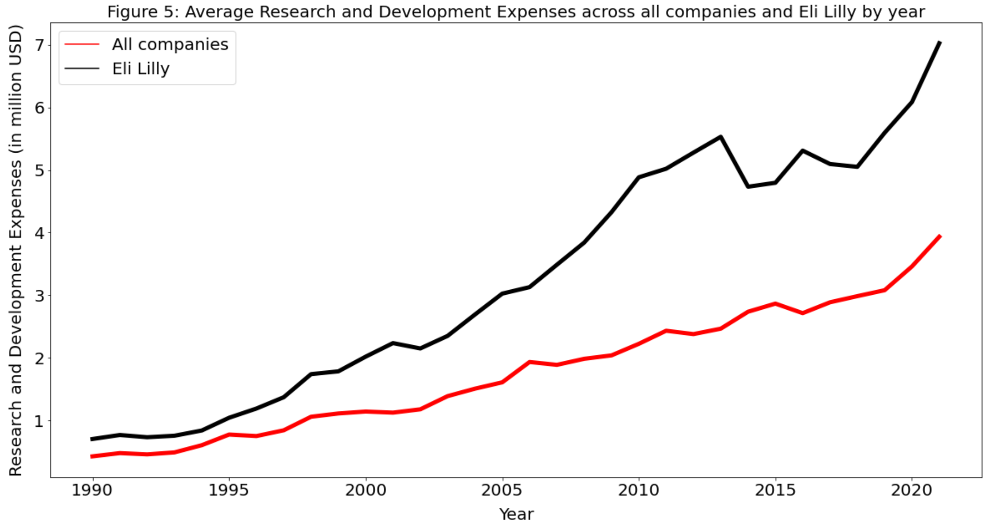
### *Total Revenue*

Based on Figure 4, it appears that Eli Lilly's total revenue has consistently been higher than the average for all companies for each of the years included in the data. In some years, the difference between Eli Lilly's total revenue and the average for all companies is relatively small, while in other years the difference is more significant. However, in no year is Eli Lilly's total revenue lower than the average for all companies. It comes close to the average in the period of 1994 - 1996.



*R&D Expenses*

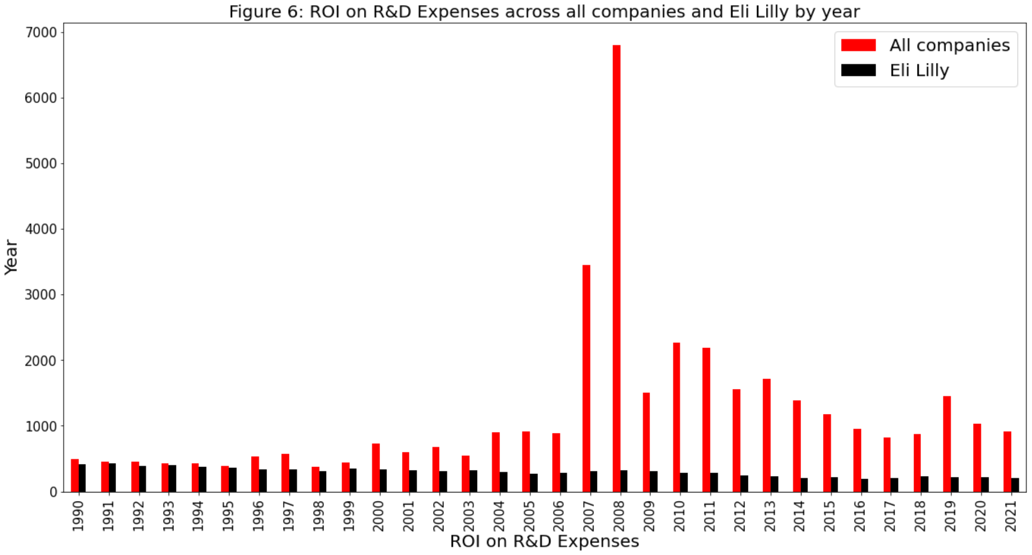
From Figure 5, we can see that in the years 1999 and onward, Eli Lilly's R&D expenses were relatively large compared to the average for all companies. In these years, Eli Lilly's expenses consistently exceeded the average. Overall, it appears that Eli Lilly had relatively large differences in R&D expenses compared to the average for all companies in recent years, but the gap was smaller in the early years.



However, it is also crucial to check whether these expenses are being spent in the right direction. ROI is checked which provides the gain from the investment {Hassanzadeh, 2019 #76}.

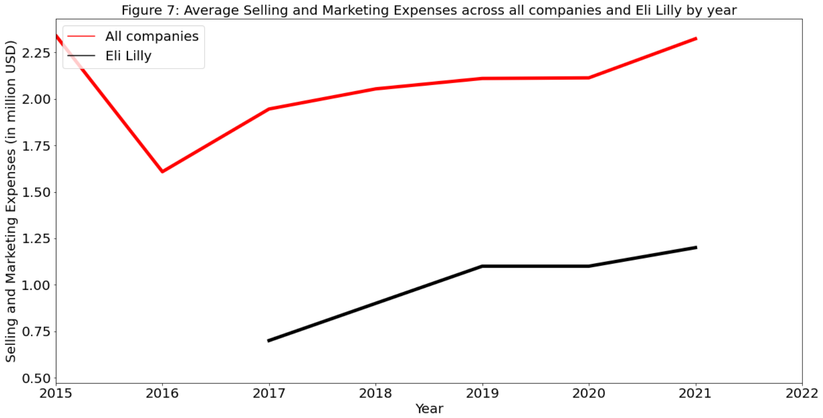
$$\text{ROI} = \frac{\text{Gain from Research Expenses} - \text{Research Expenses}}{\text{Research Expenses}} * 100$$

Upon plotting the ROI from investment in research in Eli Lilly against other companies, we get Figure 6.



Eli Lilly is clearly not directing the R&D expenditure in the correct direction as compared to the average company. Eli Lilly stands below average in ROI on multiple years, especially 2007 and 2008 by a vast margin. 2008 was also a year of negative net income for Eli Lilly which could be the reason for lower expenditure on R&D.

*Selling and Marketing Expenses*



Eli Lilly has only displayed Selling and Marketing Expenses from 2017 onwards. From Figure 7, Eli Lilly's selling and marketing expenses have been lower than the average for all companies. The difference between Eli Lilly's expenses and the average has been around 40% to 50% in these years.

*Patents*

From Figure 8, Eli Lilly's number of patents have been lower than the average for all companies in most years, except for a few years in the early 2000s and the period from 2014 to 2021.

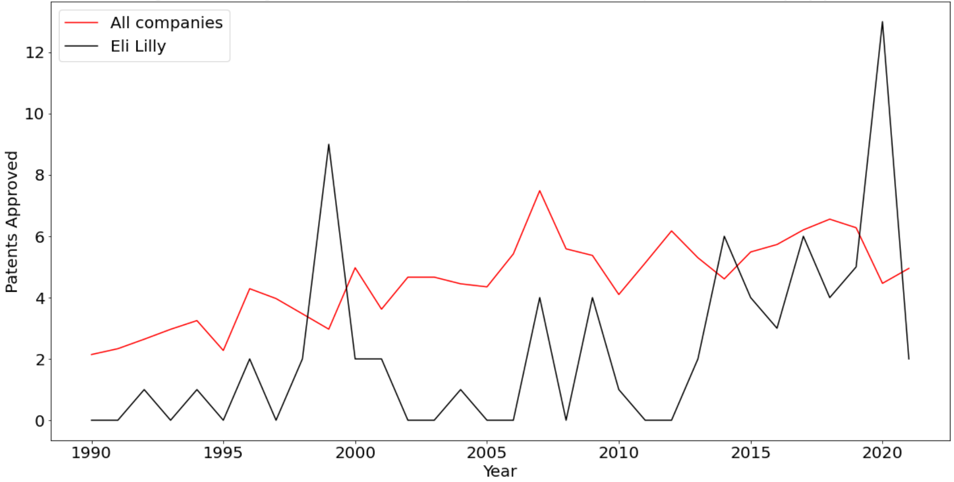


Figure 8: Average Number of Patents Approved across all companies and Eli Lilly by year

*Correlation between the variables*

In Figure 9, the table shows the correlation between R&D expenses, selling and marketing expenses, gross profit, net income, total revenue, and patents.

From the table, there is a strong positive correlation between R&D expenses and gross profit (0.95), net income (0.78), and total revenue (0.91). This means that as R&D expenses increase, these other variables also tend to increase.

There is a moderate positive correlation between R&D expenses and selling and marketing expenses (0.50). This means that as R&D expenses increase, selling and marketing expenses also tend to increase, although the relationship is not as strong as the other variables.

There is a small negative correlation between R&D expenses and patents (-0.07). This means that as R&D expenses increase, the number of patents tends to decrease, although the relationship is not very strong.

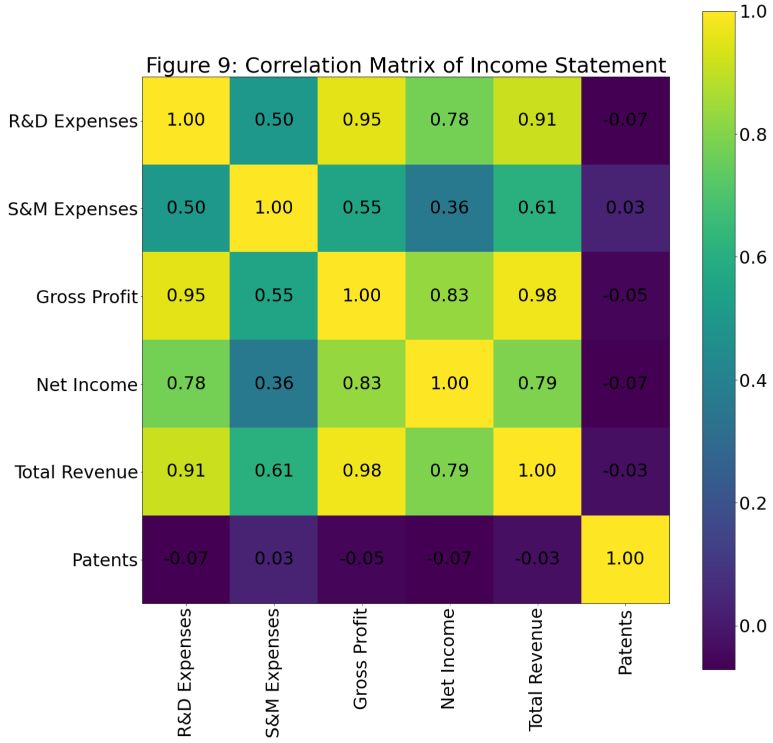


Figure 9: Correlation Matrix of Income Statement

| | R&D Expenses | S&M Expenses | Gross Profit | Net Income | Total Revenue | Patents |
|---|---|---|---|---|---|---|
| R&D Expenses | 1.00 | 0.50 | 0.95 | 0.78 | 0.91 | -0.07 |
| S&M Expenses | 0.50 | 1.00 | 0.55 | 0.36 | 0.61 | 0.03 |
| Gross Profit | 0.95 | 0.55 | 1.00 | 0.83 | 0.98 | -0.05 |
| Net Income | 0.78 | 0.36 | 0.83 | 1.00 | 0.79 | -0.07 |
| Total Revenue | 0.91 | 0.61 | 0.98 | 0.79 | 1.00 | -0.03 |
| Patents | -0.07 | 0.03 | -0.05 | -0.07 | -0.03 | 1.00 |

*Checking Causal Relationship*

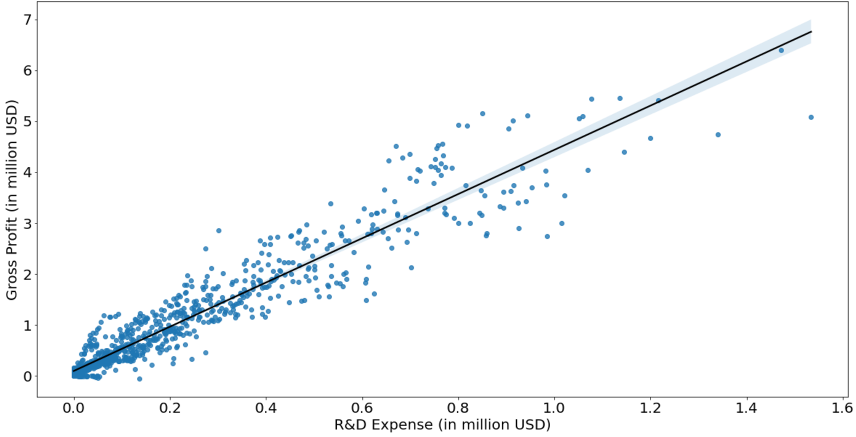
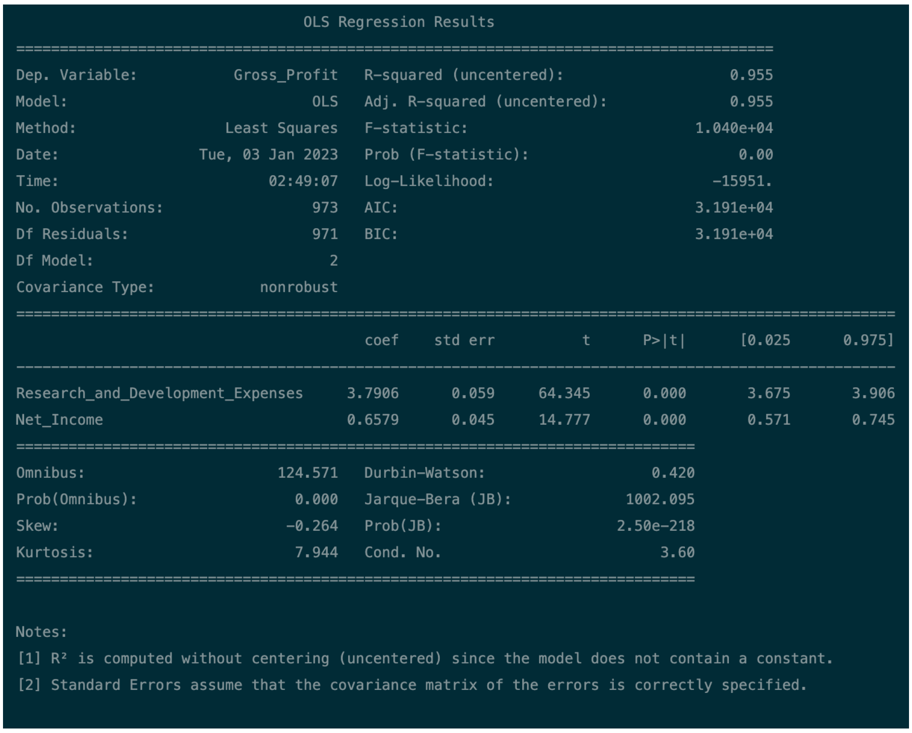


Figure 10: Regression between R&D Expense and Gross Profit

In Figure 10, the regression between R&D expenses and gross profit is being checked visually to understand whether a causal relationship exists. The two variables show a high degree of correlation and upon checking the scatterplot there indeed seems to be a positive causal relationship between them. The dots do not show much of a scatter from the line and show clustering around the regression line. However, it is not possible to use only these to model the relationship as pharmaceutical companies require a control variable to help control for size of the company. Hence, net income of the company can be used as a control variable in this situation.


```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:           Gross_Profit   R-squared (uncentered):                   0.955
Model:                            OLS   Adj. R-squared (uncentered):              0.955
Method:                 Least Squares   F-statistic:                          1.040e+04
Date:                Tue, 03 Jan 2023   Prob (F-statistic):                        0.00
Time:                        02:49:07   Log-Likelihood:                         -15951.
No. Observations:                 973   AIC:                                  3.191e+04
Df Residuals:                     971   BIC:                                  3.191e+04
Df Model:                           2
Covariance Type:            nonrobust
=====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
Research_and_Development_Expenses    3.7906      0.059     64.345      0.000       3.675       3.906
Net_Income                           0.6579      0.045     14.777      0.000       0.571       0.745
==============================================================================
Omnibus:                      124.571   Durbin-Watson:                   0.420
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1002.095
Skew:                          -0.264   Prob(JB):                    2.50e-218
Kurtosis:                       7.944   Cond. No.                         3.60
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 11: Model relationship between Gross Profit and R&D expenses while controlling for Net Income

In Figure 11, the R-squared value is 0.955, which means that 95.5% of the variance in Gross Profit is explained by the two independent variables. The F-statistic is very high (10,400), and the p-value is very small (0.00), which suggests that the model is significantly better than a model with no independent variables. The coefficient for R&D Expenses is 3.7906, which means that for every 1 unit increase in R&D Expenses, 3.7906 unit increase in Gross Profit is expected. Similarly, the coefficient for Net Income is 0.6579, which means that for every 1 unit increase in Net Income, a 0.6579 unit increase in Gross Profit is expected.

Overall, this model appears to be a strong fit for the data, as evidenced by the high R-squared value and significant F-statistic. The coefficients for the independent variables also suggest a meaningful relationship between the variables and the dependent variable. The Omnibus, Skew, and Kurtosis values are all relatively high, which may indicate that the residuals are not normally distributed. However, upon closer analysis of the residuals for skew in Figure 12, the distribution of the residuals is acceptable and there is a relatively normal distribution.

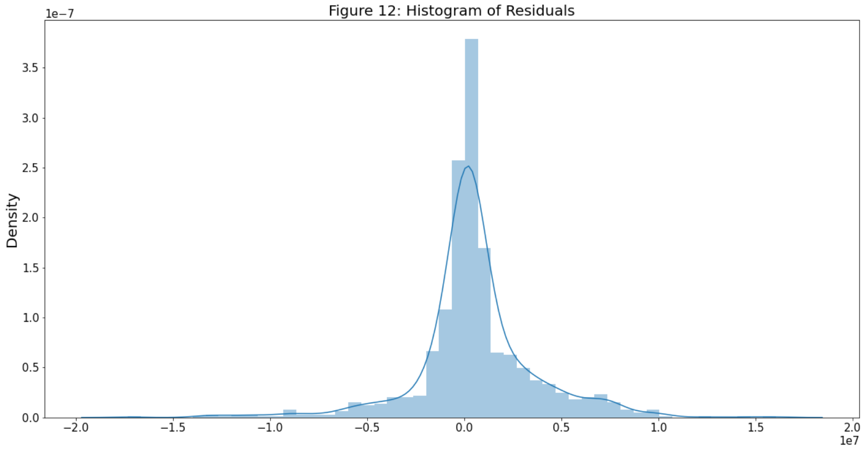


Figure 12: Histogram of Residuals

Finally, the model is checked for multicollinearity by checking VIF values.

| Feature | VIF |
|---|---|
| Research_and_Development_Expenses | 3.486411 |
| Net_Income | 3.486411 |

The VIF values of ~3.5 indicate that there is moderate multicollinearity present in the model. It is generally recommended to keep the VIF values below 5 to avoid multicollinearity. However, the presence of multicollinearity does not necessarily mean that the model is invalid. The coefficient estimates may be less reliable, and the standard errors of the coefficients may be larger.

*Predictive Model*

As is visible from the causal relationship exhibited above, it is possible to grow gross profit by investing more in the research and development area. After multiple trial and errors accounting for the lowest mean squared error and highest R-squared value, Random forest model was chosen.

| **Model** | **MSE** | **R2** |
|---|---|---|
| Linear | 11054141995127.854 | 0.8995590082051271 |
| Lasso | 11054141995128.137 | 0.8995590082051246 |
| Ridge | 11054141995127.822 | 0.8995590082051275 |
| Random Forest | 8283254184582.516 | 0.9247360612922111 |
| Decision Tree | 10763755211553.352 | 0.9021975428430213 |

These are the two indicators of fit of data. According to the Scikit-Learn documentation, “A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree” (Pedregosa *et al.*, 2011b). Decision trees help calculate values for multiple decisions to derive their utility. Random forests are a variant of decision trees also known as ensemble decision trees. They help improve performance and can be thought of as a nearest neighbour predictor (Ghavami, 2019). Upon checking for fitting of data visually as demonstrated in Figure 13, there seems to be a good fit of data.

Figure 13: Actual Values vs Predicted Values

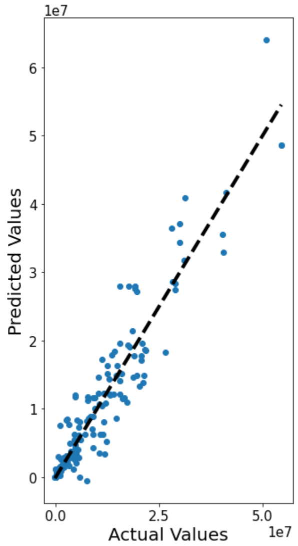
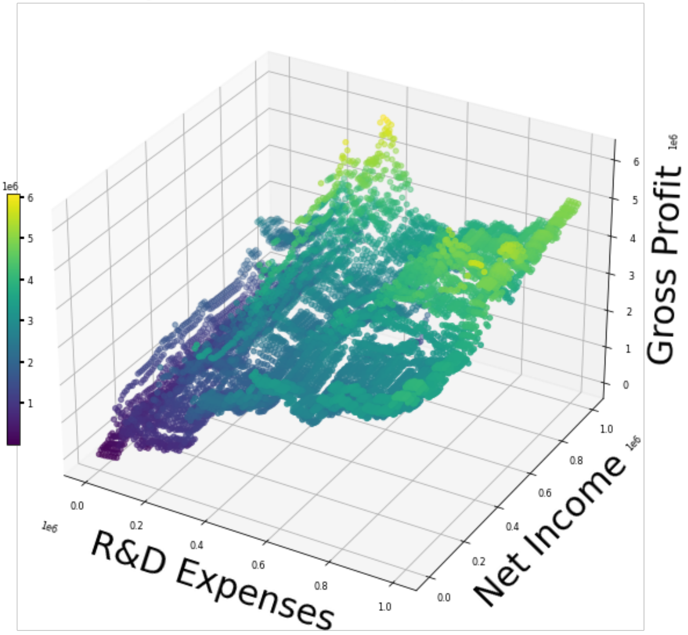


Figure 14 is a color map produced by using the predictive model. It is visible that the higher gross profit is closely related with both higher R&D expenses. The higher we go in R&D expenses, the higher the gross profit.

Figure 14: Gross Profit According to Values of Research and Development Expenses and Net Income as a Color Map



By utilizing the predictive model, it is possible to extrapolate the ideal amount of investment by Eli Lilly to match the gross profit earned by the highest earning company (Johnson & Johnson in this case). Johnson & Johnson had the highest earnings in 2021 with 21345300 ($000) dollars. Using this figure as the ideal that Eli Lilly wishes to accomplish, the required investment in research and development can be derived using the predictive model.

The ideal R&D expense for achieving the gross profit: 45746073.07
Actual R&D Expenses: 7025900
Difference between the two: 38720173.07

*Findings*

Based on the analysis and results presented above, it appears that Eli Lilly has higher gross profit, net income, and total revenue compared to the average of all companies in the data. This could be due to a variety of factors such as higher prices for its products, greater efficiency in production, and other competitive advantages. Eli Lilly's R&D expenses have consistently exceeded the average for all companies in recent years. However, upon further digging, the ROI from this expenditure is lower than average which indicates that the R&D expenses are not being utilized in the appropriate direction.

The company has had lower selling and marketing expenses compared to the average. The number of patents held by Eli Lilly has generally been lower than the average for all companies, except for a few years in the early 2000s and the period from 2014 to 2021. It is not clear what impact this has on the company's financial performance.

The correlation matrix shows that the major factors correlated to a pharmaceutical company's growth are R&D expenses, net income, and total revenue. Selling and marketing expenses are not significantly contributing towards boosting the gross profit of the company.

*Recommendations*

The difference between the ideal R&D expenses and the actual R&D expenses is approximately 55.17% between the ideal and the actual. This suggests that Eli Lilly is currently investing approximately 45.83% less in R&D than what would be ideal based on the model. Increasing investment in R&D by this amount could potentially lead to an increase in Gross Profit for Eli Lilly. However increased investment is only one side as this expense also needs to be spent in a more targeted manner which allow for greater ROI.

Increased number of patents and selling and marketing expenses do not remain the focus of the pharmaceutical industry as the number of patents are not a significant factor for determining profitability.

## Conclusions

There is a strong positive correlation between gross profit, R&D expense, and net income. This indicates that increased investment in R&D leads to higher financial performance. This is supported by the findings of the regression analysis. The random forest model, which had the highest R2 value (0.925) and the lowest MSE among all the models tested, is selected to check the ideal R&D expense for achieving the gross profit (a difference of approximately 55.17%). However, there may be other factors that contribute to the company's financial success, and further analysis is needed to understand the underlying reasons for Eli Lilly's financial performance.

*Limitations*

According to ICAEW, the company size is measured using annual turnover or the Total Revenue as one of the components. However, usage of this variable leads to high multicollinearity and over fitting of the data. To compensate for this, net income is being used in its stead. The accuracy of using net income as a control variable for company size is yet to be determined. Using number of employees is probably a better approach but due to lack of data on number of employees for most of the years of the data that is on hand, this approach was eliminated for the time being.

In this case, the MSE is quite large, which suggests that the model is not making very accurate predictions. A lower MSE indicates a better fit of the model to the data. All the models (Linear, Lasso, Ridge, Random forest, Decision tree) have a very high MSE value, which suggests that they are not fitting the data very well. This could be due to several factors, such as a lack of sufficient data, incorrect model assumptions, or the presence of multicollinearity or other sources of noise in the data.

To improve the model fit, more predictors could be added to the model, or a different type of regression model altogether. Different techniques could also be used for preprocessing the data (e.g., normalization, feature selection) to improve the quality of the input data.

Random forest is being utilised for predictive modelling however it has a major drawback wherein upon used for regression, it cannot predict beyond the range of the training data. It is also prone to over-fitting noisy data sets (Ghavami, 2019).

## References

Administration, U. F. D. (2022) 'Purple Book Database of Licensed Biological Products', Purple Book Database of Licensed Biological Products. Available at: https://purplebooksearch.fda.gov/downloads (Accessed: December 17, 2022).
Company Profiles: Eli Lilly & Company. Marketline.
Davidson, S., Stickney, C. P., Weil, R. L., Stickney, C. P. and Weil, R. L. (1979) *Financial accounting : an introduction to concepts, methods, and uses.* 2d edn. Hinsdale, Ill.: Dryden Press, p. xviii, 695 pages ; 25 cm.
Executive, P. (2022) 'Enterprise value of leading pharmaceutical companies worldwide as of 2021 (in billion U.S. dollars)'. Available at: https://www-statista-com.bris.idm.oclc.org/statistics/473326/enterprise-value-for-top-global-pharmaceutical-companies/ (Accessed: December 17, 2022).
Ghavami, P. (2019) *Big Data Analytics Methods : Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing* Berlin ;: De Gruyter. Available at: https://doi.org/10.1515/9781547401567
https://www.degruyter.com/doc/cover/9781547401567.jpg.
Harris, C. R. a. M., K. Jarrod and van der Walt, Stéfan J and Gommers, Ralf and Virtanen, Pauli and Cournapeau, David and Wieser, Eric and Taylor, Julian and Berg, Sebastian and Smith, Nathaniel J. and Kern, Robert and Picus, Matti and Hoyer, Stephan and van Kerkwijk, Marten H. and Brett, Matthew and Haldane, Allan and Fernández del Río, Jaime and Wiebe,

Mark and Peterson, Pearu and Gérard-Marchant, Pierre and Sheppard, Kevin and Reddy, Tyler and Weckesser, Warren and Abbasi, Hameer and Gohlke, Christoph and Oliphant, Travis E. (2020) 'Array programming with {NumPy}', *Nature,* 585, pp. 357–362.

Hunter, J. D. (2007) 'Matplotlib: A 2D graphics environment', *Computing in science & engineering,* 9(3), pp. 90-95.

Kimmel, P. D., Weygandt, J. J. and Kieso, D. E. (2007) *Financial accounting : tools for business decision making.* 4th edn. Hoboken, N.J: John Wiley & Sons.

Lakdawalla, D. N. (2018) 'Economics of the Pharmaceutical Industry†', *Journal of Economic Literature,* 56(2), pp. 397-449.

McKinney, W. and Others 'Data structures for statistical computing in python', *Proceedings of the 9th Python in Science Conference*, 2010, 51-56.

'Orange Book Data Files' (2022), Orange Book Data Files (compressed). Available at: https://www.fda.gov/drugs/drug-approvals-and-databases/orange-book-data-files (Accessed: December 17, 2022).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Others (2011a) 'Scikit-learn: Machine learning in Python', *Journal of machine learning research,* 12(Oct), pp. 2825-2830.

Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V., and Thirion, B. a. G., O. and Blondel, M. and Prettenhofer, P., and Weiss, R. a. D., V. and Vanderplas, J. and Passos, A. and and Cournapeau, D. a. B., M. and Perrot, M. and Duchesnay, E. (2011b) 'Scikit-learn: Machine Learning in {P}ython', *Journal of Machine Learning Research,* 12, pp. 2825--2830.

S&P Capital IQ (Firm) 2013. Current market perspectives. New York, New York: S&P Capital IQ.

Scherer, F. M. (2000) 'Chapter 25 The pharmaceutical industry', *Handbook of Health Economics*: Elsevier, pp. 1297-1336.

Seabold, S. and Perktold, J. 'statsmodels: Econometric and statistical modeling with python', *9th Python in Science Conference*, 2010.

Van Rossum, G. and Drake, F. L., Jr. (1995) *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. and SciPy, C. (2020) 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods,* 17, pp. 261-272.

Author (2017) *mwaskom/seaborn: v0.8.1 (September 2017)* (Version v0.8.1). Available at: https://doi.org/10.5281/zenodo.883859
http://dx.doi.org/10.5281/zenodo.883859.