MIT | TAKEDA PHARMACUETICALS

Milestone 5: Final Project Report

Document encapsulating Milestone 5 for 15.089 Analytics Capstone

Edoardo Italia & Denis Sai

eitalia@mit.edu | dsai@mit.edu

Table of Contents

1. Company Description, Problem Scope, and Motivation

Takeda Pharmaceuticals ("Takeda") is a Japan-based global pharmaceutical company, operating in approximately 80 countries. Focusing on six main areas of research – ranging from oncology to neuroscience – Takeda is a well-established player in the pharmaceutical space. Central to Takeda's philosophy are its investments in advanced research and development, made to drive innovation in drug development and treatment.

Clinical trial execution is a complex project. It requires detailed planning and a cross-functional collaboration as well as support from external stakeholders. The Global Development Office (GDO) is an internal organization that is the execution group of Takeda's Clinical Development Strategy. Takeda outsources non-core capabilities to Contract Research Organizations (CROs) who play a key role in the operational execution of clinical trials.

GDO has created a 3-year strategy and vision to further strengthen analytical capabilities to support an efficient execution of clinical trials that is focused on providing "Predictable Delivery" of clinical trials. Takeda aims to build an operational capability to predict the delivery of segments of clinical trials in the future with a high certainty. Their desire is "to transition from retrospective performance analysis to proactive measurements which enable them to predict performance and course-correct as needed".

AI-based methodologies can be used to derive hidden insights, used then to administer actionable steps aimed at improving operational efficiencies and predicting delivery times. To this end, Takeda has tasked us to explore and develop analytics-based solutions for the following topics:

1) **Benchmarking clinical trial operational performance** metrics by expanding current analytical capabilities

2) **Classifying** "good" and "bad" **performances** across therapeutic areas (TAs) and Contract Research Organizations (CROs)

3) **Predicting** -and, potentially, optimizing- any **variance** between planned and actual clinical trial developments

Problem Scope

After several additional conversations, we have decided to focus our attention on the period of clinical trials between dates of Last Subject Out (LSO) and Database Lock (DBL) – referred to as Close Out. Here one of the problems that can occur is Database Unlocks after the initial Database Lock. Below is a breakdown of each of the terms:

- **Last Subject Out (LSO)** – the date when we received relevant information from the last patient enrolled in a clinical trial alongside all metrics and parameters received during the clinical trial

- **Database Lock (DBL)** – date when we all CRO-derived information for a given trial is processed and prepared for statistical analysis

- **Close Out stage –** time between LSO and DBL

- **Database Unlocks –** event that happens when, after DBL, we are required to open a data review to either add or recollect data critical to the statistical analysis (drug patient response) after DBL.

The scope of the Capstone Project was honed on the analysis of the performance of specific milestones in clinical trials:

1) **Close Out stage of Clinical Trials:** This stage effectively represents the final node of a given clinical trial. It is critical to complete this step accurately and efficiently to ensure that any clinical insights or analyses completed are correct and up to internal standards.

2) **Analysis of Unlocks of Database between LSO and DBL:** Analyze number of unlocks for each clinical trial to generate insights for future clinical trials whether unlocks are going to happen. It is important to predict whether unlocks are going to happen in advance since Takeda can increase the number of quality checks for trials that are "highly likely" to be problematic.

Prioritized in the given order, these two stages of the clinical trials have been identified for us to dive into further. To this end, deliverables pertaining to any insights generated, machine-learning and/or data processing scripts, have been highlighted as particularly useful.

Expected Goals

Takeda is looking to understand key drivers (e.g. Therapeutic Area, Phase, CRO, etc.) that are impacting performance in the Close Out Stage (LSO to DBL) and impacting number of unlocks for a trial, and understand how AI/ML may help to identify early risk indicators to projected timelines. Moving from this, we identified several goals that we were looking to achieve by the end of August.

For Close Out analysis:

1. Understand Takeda's current position as compared to industry performance over time

2. Compare Takeda's performance in the Close Out stage between Pre-COVID and during COVID time by Therapeutic Area, CRO and Phase

3. Find key drivers of current Close Out performance and provide, if feasible, a "Predictable Delivery" tool to predict Close Out time for future clinical trials

For Unlocks analysis:

1. Find available data on unlocks of clinical trials for the last several years

2. Analyze distributions of number of unlocks per Therapeutic Area, CRO, Phase of clinical trials

By the end of our internship, we managed to successfully reach all expected goals. The following sections will thus elaborate on our work.

Motivation

Clinical trials represent one of the most crucial components of a new therapy's pipeline. Subdivided into three main phases (I, II, III) with a potential Phase IV after commercial launch, clinical trials are conducted to evaluate the safety, toxicity, and effectiveness of any drug or therapy created. The laborious processes in clinical development often incur on the order of hundreds of millions of sunk costs for pharmaceutical companies carrying out this research. The investment required represents the importance and fragility of this process, which involves countless stakeholders (patients, partnering clinics, CROs, etc.) and last numerous years.

Optimizing the performance of these trials is thus key if companies such as Takeda are to pursue increasing amounts of research. A successful implementation of the "Predictable Delivery" tools would ultimately not only improve the efficiency of any capital deployed, but also potentially reduce a drug's time-to-market while increasing certainty in the conclusions obtained from any studies.

In addition to the elements described above, we have been asked to confirm the following points resulting from internal analysis conducted prior to our arrival:

1. Slight improvements in Close Out timings were noted in 2018-2019, but 2020 has seen increased cycle times

2. Data indicates there is some variability in the delivery of cycle times over the past years

3. There is a high prevalence of data unlocks

From our initial analysis, we found that the same trend of inflated cycle times is observed in 2021, shown in *Figure 1*.

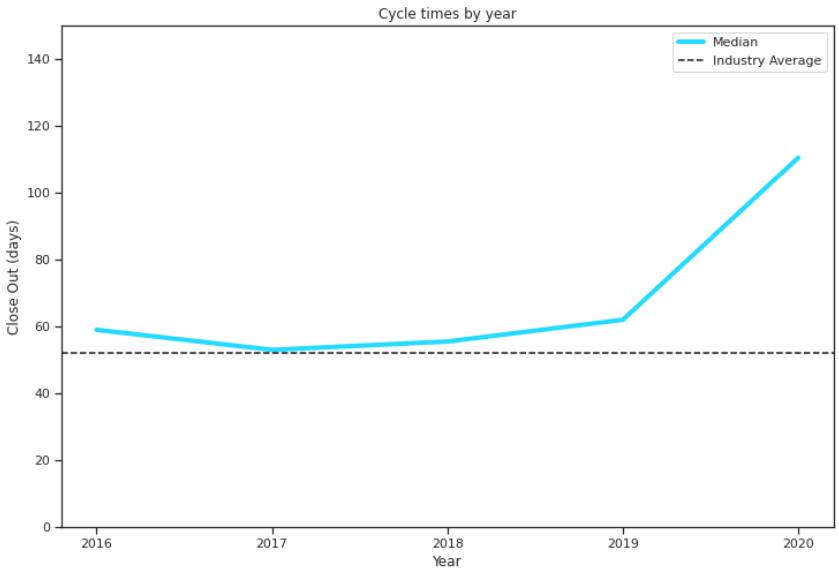


Figure 1 Trailing cycle times medians, 2016 - 2020

We can observe that while industry average close out time is stable at ~52 days for a Close Out Stage, Takeda observes increased cycle time from ~60 days in 2019 to ~110 days in 2020.

As each additional month of clinical trials increases the cost of tests by millions of dollars and lengthens the time to market for a potentially successful drug, identifying the parameters affecting the Close Out time is currently among Takeda's top priorities.

2. Data Breakdown

Data on Close Out

Importance

Understanding close out cycle times for the end of a clinical trial is essential for researchers to manage the post-study statistical analyses. For processes like clinical trials, where it is crucial to complete all steps on time, having a view on the causes of delay or speed-ups provides an edge in project planning. The data tables our work dives into provides this view for trial close out.

Data Collection

The operational performance of a clinical trial is captured on an ongoing basis, with milestones and associated timelines sequentially included in Takeda's data lakes through project snapshots. During these snapshots, the status of a project is updated, with dates, task owners, as well as trial features refreshed all to describe the current project standing, thereby providing a view on the expected versus actual operational performance. In this regard, data

management is almost entirely automated, requiring only updated inputs from the relevant task owners for a given trial. Once collected, the data is stored in the datahub and either maintained on the servers or processed and presented in internal dashboards.

Structure

Analyzed operational data has come primarily from two sources: raw relational databases, and an internal Clinical Analytics Report (CAR). The former was engineered by us, whereas the CAR dataset was created by Takeda's internal team of data experts. Following some additional feature engineering, the data used for close-out cycle time analyses had the following features:

- Trial ID (alphanumeric string)
    - Represents the unique identifier code for a specific trial. A single candidate drug may share the same code root, but no observation shares the same unique code
- LSO (date)
    - Represents the date of completion for the Last Subject Out milestone
- DBL (date)
    - Represents the date of completion for the Database Lock milestone
- Protocol Phase (categorical)
    - Represents the phase of the trial, broken into four groups: I, II, III, and IV
- Therapeutic Area (categorical)
    - Represents the therapeutic area (TA) of a trial. Broken into five groups: Oncology, Gastroenterology, Neuroscience, Rare Diseases, and non-core TAs.
- Contract Research Organization (categorical)
    - Represents the research partners (CRO) pharmaceutical companies work with during clinical trials. Broken into six groups: IQVIA, PRA, PPD, Celerion, ICON, and other CROs.
- Planned Enrollment (numeric)
    - Represents the planned number of enrolled patients
- Actual Enrollment (numeric)
    - Represents the actual number of enrolled patients, after patient screening
- Percentage Enrollment Completed (numeric)
    - Represents the ratio of actual enrollment to planned enrollment
- Planned Sites (numeric)
    - Represents the planned number of clinical sites the be used in each trial
- Active Sites (numeric)
    - Represents the number of clinical sites currently being used in each trial
- Number of LSO (numeric)
    - An engineered feature representing the number of times the LSO date was changed
- Number of DBL (numeric)
    - An engineered feature representing the number of times the DBL date was changed
- Synopsis to Protocol Approval (numeric)

    - Represents the number of calendar days it took to move from the *Synopsis* to *Protocol Approval* milestones
- First Subject Selected to First Subject In (numeric)
    - Represents the number of calendar days it took to move from the *First Subject Selected* to *First Subject In* milestones, representing the start of the trial
- Last Subject In to Last Subject Out (numeric)
    - Represents the number of calendar days it took to move from the *Last Subject In* to *Last Subject Out* milestones
- COVID19 (binary):
    - A binary feature representing whether this trial occurred during COVID-19 (taken to be after March 15th 2020)
- Close Out (numeric):
    - Represents the number of calendar days between the observed dates of completion for the *Last Subject Out* and *Database Lock* milestones. This is our dependent variable.

The observations track the operational cycle times from 2012 till the most recent snapshot and are updated monthly. Important to note is that Takeda has undergone numerous large acquisitions in this time frame, including ARIAD Pharmaceuticals (completed in 2017) and Shire (completed in 2019). The trial observations are thus inclusive of legacy data from the acquired parties. Finally, while the fully raw dataset contains roughly 400 observations, upon further processing and feature engineering, the dataset contained only roughly 300 observations.

Data on Unlocks

Importance

Clinical trials are continuously ongoing processes, with research often being conducted on the same candidate drug during the periods of statistical analysis. Database unlocks slow down both clinical research and post-study analyses as these require data reviews to ensure that the correctly structured information is obtained. These slowdowns incur significant time and financial expenses for pharmaceutical companies – leading these to understand what the cause of unlocks may be. Further, a database unlock can be used as a proxy to gauge the quality of the data collected by research partners, critical information that can be used in the planning of future trials.

Data Collection

The occurrence of database unlocks has come under Takeda's spotlight only in recent years. As such, automatic data unlocks flags or updates are yet to be integrated within Takeda's data pipeline. Currently, the unlock occurrences are manually recorded by task owners, who -for a given trial- report the data lock, unlock, and relock dates, as well as a short summary justifying the unlock. This dataset is not stored in either of the servers we have been given access to and started tracking unlocks only in 2019.

Structure

Data pertaining to data unlocks during clinical research is collected manually and structured into the categories below:

- Study ID (alphanumeric string)
  - Equivalent to the *Trial ID* feature in the previous dataset
- TAU/BU (categorical)
  - Equivalent to the *Therapeutic Area* feature in the previous dataset
- CRO (categorical)
  - Equivalent to the *Contract Research Organization* feature in the previous dataset
- Type (categorical)
  - Represents the type of analysis conducted. Broken into two categories: final and interim.
- Lock date (date)
  - Represents the date when the first database lock was conducted
- Unlock date (date)
  - Represents the date when the database was unlocked
- Relock date (date)
  - Represents the date when the database was re-locked, following unlock
- Comments (string)
  - Represents any comments explaining the reasons for database unlock

Each observation here represents an unlock for a specific trial. To maximize the number of features describing the unlocks by trial, the unlock data above was joined with the processed data for close-out analysis.

3. Analytical Methods

Tools Used

Resulting from the large data wrangling effort and desired visualizations and predictions, we have been fortunate enough to make use of a range of analytical tools. The primary languages we are working with are SQL (through Takeda's Microsoft-SQL servers) and Python (through Takeda's JupyterHub), with the latter requiring us to leverage numerous packages and libraries. These include, but are not limited to, pandas, numpy, seaborn, scipy, and scikit-learn.

To maximize the likelihood of correctly predicting cycle times and database unlocks, we decided to leverage a series of machine learning approaches. Ranging from more to less interpretable models, we attempted predicting outcomes through regression, binary classification, and multi-class classification-based frameworks. The *Machine Learning Techniques* section below provides a deeper dive into the reasoning behind the selected approaches.

Visualization Techniques

To provide a holistic overview of the available data, we took a sequential approach. Starting first with a generic exploratory data analysis (EDA) to understand the overall cycle-time performance as well as the distribution of trials, we then dived into more detailed breakdowns to cross compare trial categories. Of importance to Takeda is understanding the variation in performances across the CROs used, phases researched, and therapeutic areas studies. This is to be combined with a desire to understand where, if at all, the COVID-19 pandemic has had an impact on operational performance. To complete this task, we broke down the close-out performances, as well database unlock prevalence, into the specified categories above.

An additional crucial element to be analyzed is that of understanding what could lead to a top or bottom quartile operational performance. As part of an industry-wide push to accelerate clinical trials, pharmaceutical companies constantly compare themselves to industry benchmark, thereby aiming to pursue top-quartile performance. In satisfying this demand, we delved into evaluating the presence of outliers across the three trial features specified above.

Machine Learning Techniques

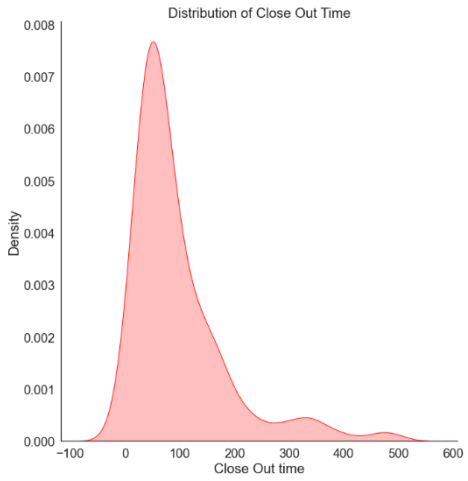


Figure 2 Density Distribution of Close Out Times

As our main goal was to characterize and evaluate drivers of variability in performance and delays of close out, we tried two main machine learning frameworks to solve the main problem:

1. **Regression Approach** – predict actual time to conduct a Close Out stage (between Last Subject Out and Database Lock). For this analysis, we used several interpretable and black-box approaches:

    a. *Linear Regression* – to try to model Close Out assuming linear relationships between dependent variable and set of independent variables

b. *L1, L2 Linear Regressions* – to use Linear Regression but with different types of regularizations

c. *Poisson Regression* – since we have a non-normal distribution of our dependent variable (as shown in *Figure 2*) we try to model it with Poisson Regression that takes this information into account

d. *Regression Decision Trees* – use a model with tree-like structure of decisions and their possible consequences on Close Out time prediction

e. *Gradient Boosting* – an ensemble model that uses week decision trees to consecutively update dependent variable by applying differentiable loss function. From academic research, on average, this method outperforms random forest and interpretable methods, but for each specific dataset it is better to try a wide range of models to identify the best one

2. **Classification Approach** – instead of predicting actual time for a specific stage (close-out), we predicted a binary variable indicating whether we achieve actual Database Lock, DBL*, on-time* with the initial internal expected Database Lock or not. In our model 1 indicates that we achieved Close Out before we expected and 0 indicates that we did not achieve Close Out on-time. For this case, we used the following set of models:

   a. *Logistic Regression* – an interpretable linear model that predicts a probability of an assignment to class 1. This model is also based on an assumption that there is a linear relationship between log odds and set of independent variables

   b. *Classification Decision Tree* – a model with a tree-like structure of decisions and their possible consequences. Close by structure to Regression Decision Trees with a difference in loss function and a way to make a prediction in the end

   c. *Random Forest* – one of the ensemble models that averages multiple deep decision trees trained on a different randomized subset of features and bootstrapped subset of observations.

   d. *Gradient Boosting* – ideas is the same as for a regression approach

   e. *Categorical Boosting (CatBoost)* – gradient boosting library that successfully handles categorical features and outperforms existing publicly available implementations of gradient boosting in terms of quality

Machine Learning Data Preparation and Splitting

We use two main steps in preprocessing our data – handling NAs and preprocessing categorical features:

- *Handling NAs* – we replaced NA values with median values for all columns with NA values expect for Lock Duration – since for unlock duration we had an additional information that before 2019 unlocks where not a big problem for a company, so all NA values would be more relevant to replace with 0 instead of a median value.
- *Preprocessing categorical features* – we applied one-hot-encoding to CRO, Protocol Phase, Therapeutic Area. We ended up with 45 processed features that were ready for machine learning analysis

After processing the data, we needed to split the data. We decided to split data into train test split by time as shown on *Figure 3*. In the end, we had 182 observations in train set, 73 observations in evaluation set, 58 observations in test set.

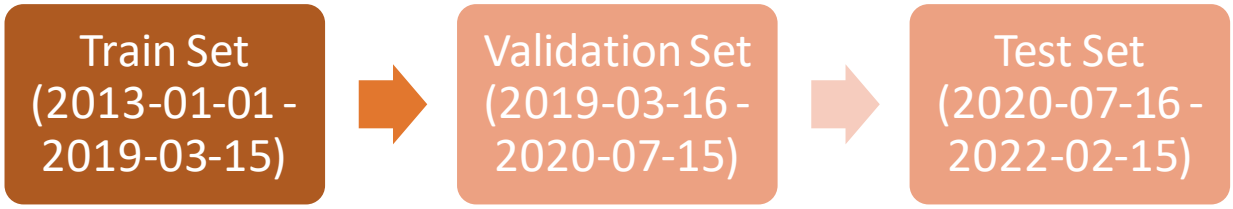


Figure 3 Train-Validation-Test Split

4. Analytical Methods and Results

Preliminary Exploratory Data Analysis

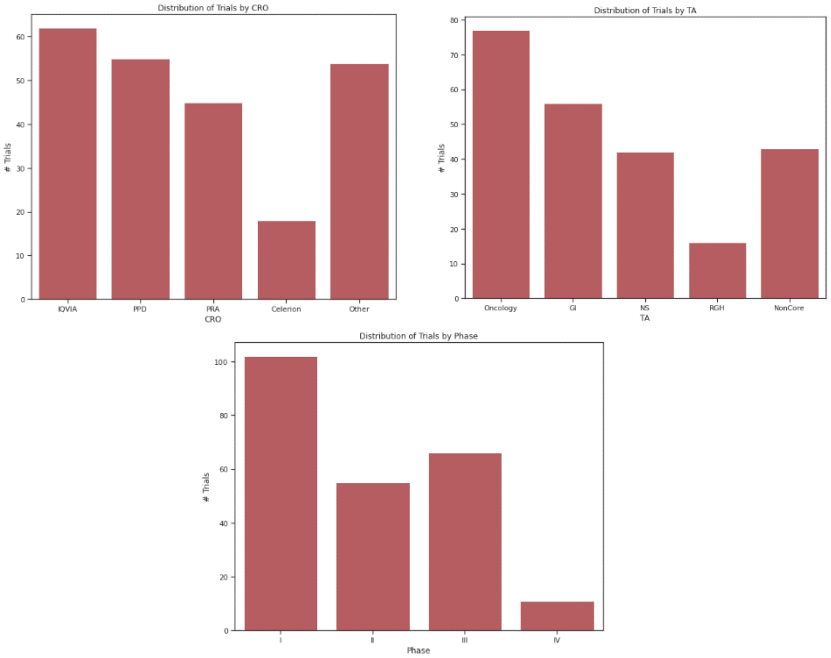


Figure 4 Distributions of Trials by Contract Research Organization (left), Therapeutic Area (right), and Phase (bottom)

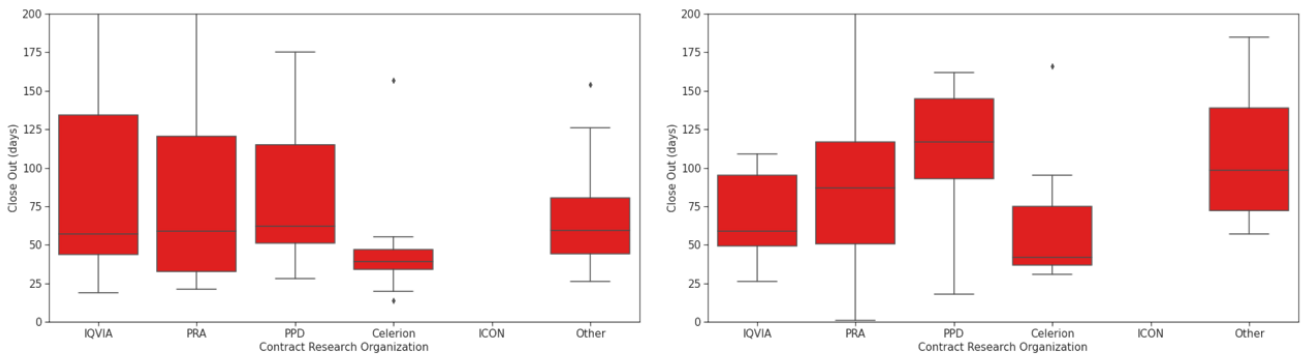
To provide context to our work, *Figure 4* provides a distribution of clinical trials across the three key categories defining a trial: Contract Research Organization (CRO), Therapeutic Area (TA), and Phase. As per the visualizations above, the three most important CRO partners to for Takeda are IQVIA, PPD, and PRA, having managed most of Takeda's historical trials. Oncology further appears to be an area of focus, with the highest number of trials having been conducted for this TA between 2012 and 2021. Finally, and resulting from the success rates across the various clinical trial phases, Phase I holds the largest share of trials conducted over the past eight years – with this share following a downwards trend as the trial phase is increased.

Given the above, insights and takeaways pertaining to the most represented categories are of prioritized importance to for Takeda.

Exploratory Data Analysis: Close Out Time

Group Variance of Close Out Times

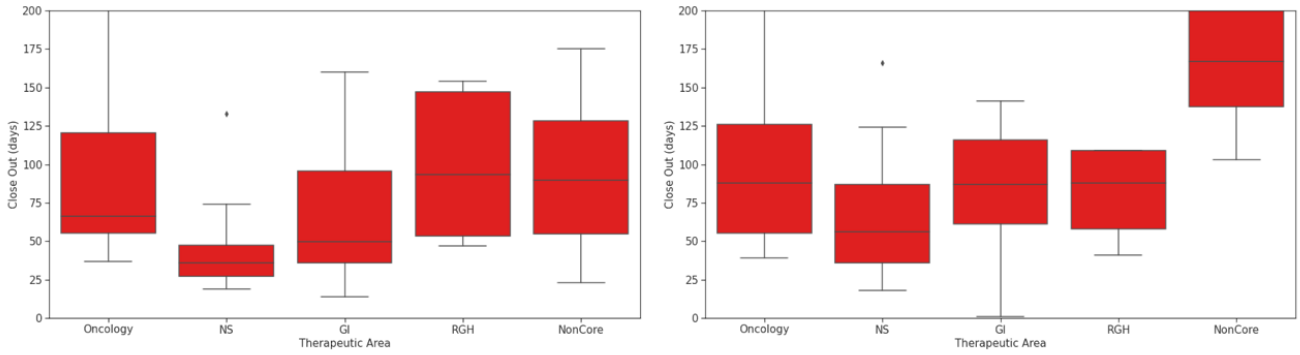
Observing the variation in close out time performance across specified groups required us to visualize the relevant data with boxplots shown below. Resulting from a continuous comparison to industry benchmarks, the analyses were conducted for the same 3-year time windows that industry data lenders provide. Specifically, we generated plots for the 2016-2018, 2017-2019- and 2018–2020-year windows. Furthermore, a comparison between pre-COVID and COVID-19 periods, within the 2018-2021 window, was conducted. For simplicity and ease of read, the pre-COVID and COVID-19 plots only are shown below. For the 2018-2021 window, the pre-COVID period contains trials whose database locks were completed between January 1st, 2018 and March 15th, 2020. March 15th, 2020 is used as the cutoff date for the beginning of the COVID-19 pandemic to reflect the enaction of national lockdowns across the world. The breakdowns were performed across Contract Research Organization (CRO), Therapeutic Area (TA), and Trial Phase.



# Trials	28	31	21	11	0	22	11	15	11	7	0	8

Figure 5 Distribution of close out times across CROs before the start of the COVID-19 pandemic (left) and after (right)

Figure 5 provides a breakdown of close times by CRO, with the number of observations within each box annotated below. The absence of ICON-managed trials pertains to the window used. It appears that from 2018 onwards, ICON did not engage in any new trials. Evident from the figure above, as for the following figures, there is great variance in close out time across CROs, especially for the three main CROs: IQVIA, PRA, and PPD. Notably, Celerion demonstrates considerably more consistent performances, albeit their expertise is focused on Phase I Oncology – thereby reducing the number of trials they engage in. *Figure 5* further acknowledges the findings in the initial EDA: the pandemic has indeed affected operational performance. While close out variance appears to concentrate during the pandemic, PRA and PPD posted dramatically higher cycle teams.



# Trials	27	22	38	6	20	17	13	13	5	4

Figure 6 Distribution of close out times across Therapeutic Areas (TAs) before the start of the COVID-19 pandemic (left) and after (right). Neuroscience (NS), Gastroenterology (GI), and the Rare Diseases (RGH) unit have been given their acronyms. NonCore represents TAs that are non-core to Takeda's business.

We notice an analogous breakdown in *Figure 6* with a focus on the trial Therapeutic Area. The impact of COVID appears more severe in this breakdown, with all TAs -except for RGH- yielding longer cycle times during the pandemic. The neuroscience division, despite suffering from COVID-related slowdowns does remain the best performing division, both in terms of median cycle time and associated variance. Notably, in a similar fashion to *Figure 5*, the variances also tend to decrease during the pandemic.

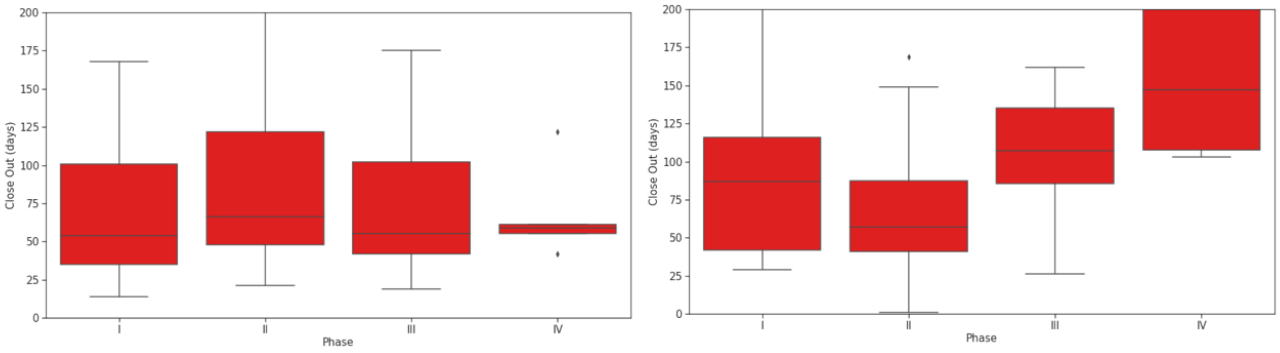


Figure 7 Distribution of close out times across Phases before the start of the COVID-19 pandemic (left) and after (right)

Finally, we can observe the breakdown by trial phase in *Figure 7*. The severe disruption is also noticed across all phases, with Phase IV – pertaining to studies conducted after the commercialization of a drug – showing the most significant increase in median close out time. Despite the increase in time, Phase IV is the least frequent of phases researched in both pre-COVID and COVID periods. The reduction in time variance is not as substantial across phases as it is across CROs and TAs; Phases II and III might have more concentrated times, though Phase I's times remain virtually unchanged. While causality cannot be inferred, there does appear to be a positive correlation between the size of the trial and the slowdown caused by the pandemic. This may be due to the restrictions imposed on interactions between people throughout the lockdowns.

Following the positive feedback received upon the presentation of these results to our colleagues at Takeda, our hope is that the data analysis above provides preliminary insights on which trials may perform more reliably than others, as well as what the expectations for the project time of a certain trial may be. Visualizing the differences in performances is the first step in understanding where to pinpoint an organization's resources, and we believe the elements above can help Takeda's GDO better focus their efforts in speeding up clinical research.

Top and Bottom Quartile Performances

Takeda's vision to accelerate clinical trial cycle times includes increasing the occurrence of top quartile performances during the close out stage. As per industry data, a top quartile close out performance is one that takes less than 33 calendar days. Ideally, there are to be no further data unlocks and such that the data is finalized for statistical analysis. Conversely, a bottom quartile performance is one with the close out cycle time taking over 82 days.

To analyze the occurrence of top and bottom quartile performances, we use the close out data set and labelled the best and worst performing trials for the 2012-2021 period. Reported below is the prevalence of these trials by CRO, Therapeutic Area, and Phase. The prevalence is obtained by dividing the number of top (or bottom) quartile performances by the total number of trials observed in the specified category.

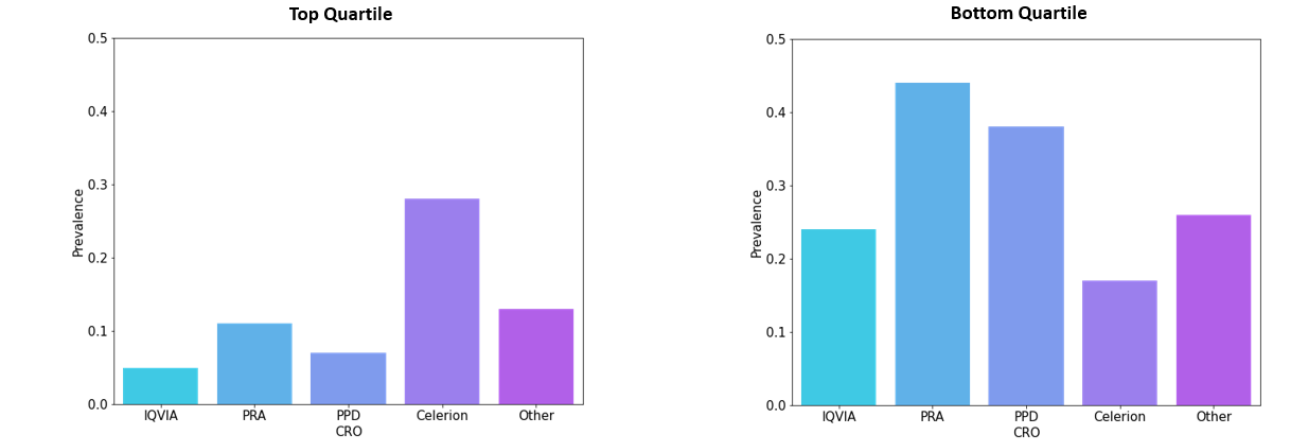


Figure 8 Prevalence of top quartile (left) and bottom quartile (right) trial performance by CRO

Figure 8 shows the prevalence of top and bottom quartile performances across CROs. Selecting an adequate research partner is a delicate process for pharmaceutical companies as these must triangulate between operational expertise, reliability, and cost. As observed, ICON and Celerion's performances standout, with these showing the highest prevalence of top quartile trials, as well as the lowest prevalence of bottom quartile trials. Importantly, these two CROs have also engaged in the fewest number of trials. Thus, while the advantage is present, it is not possible to conclude whether this would scale with additional trials. Additionally, IQVIA, PRA, and PPD, despite being the most frequent partners, all appear to have high rates of bottom quartile performances.

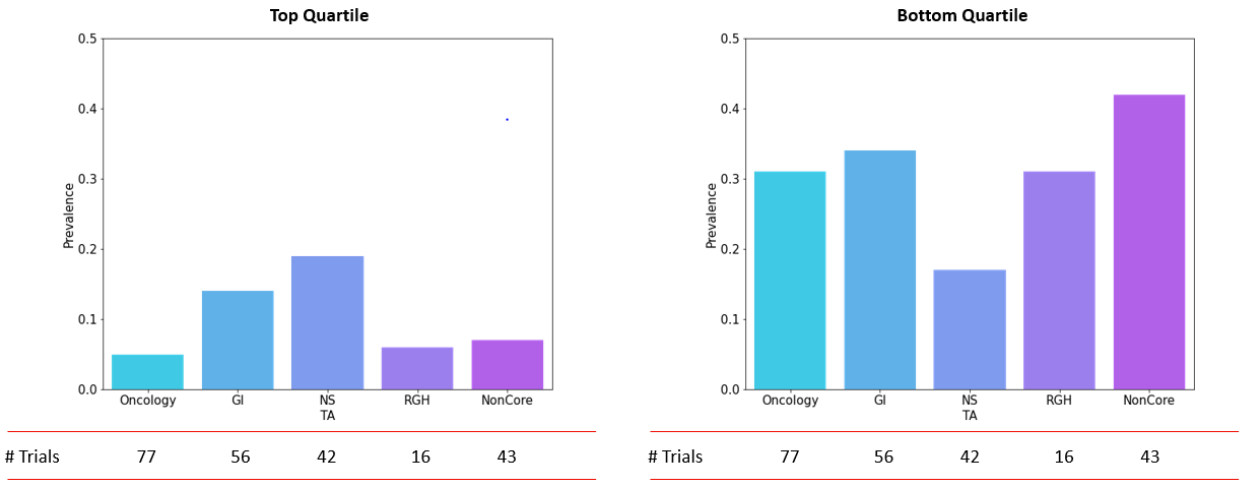


Figure 9 Prevalence of top quartile (left) and bottom quartile (right) trial performance by TA

The breakdown of quartile prevalence across Therapeutic Areas is shown in *Figure 9.* Notably, while the Neuroscience unit appears to be the strongest performer -confirming the initial results from the boxplot analysis above- there is a generally low prevalence of top quartile performance across all TAs. In fact, high rates of bottom quartile performances are found instead. The highest rates – above 40% - are observed within the Rare Diseases (RGH) and non-core units, which combine to account for 30% of trials conducted at Takeda.

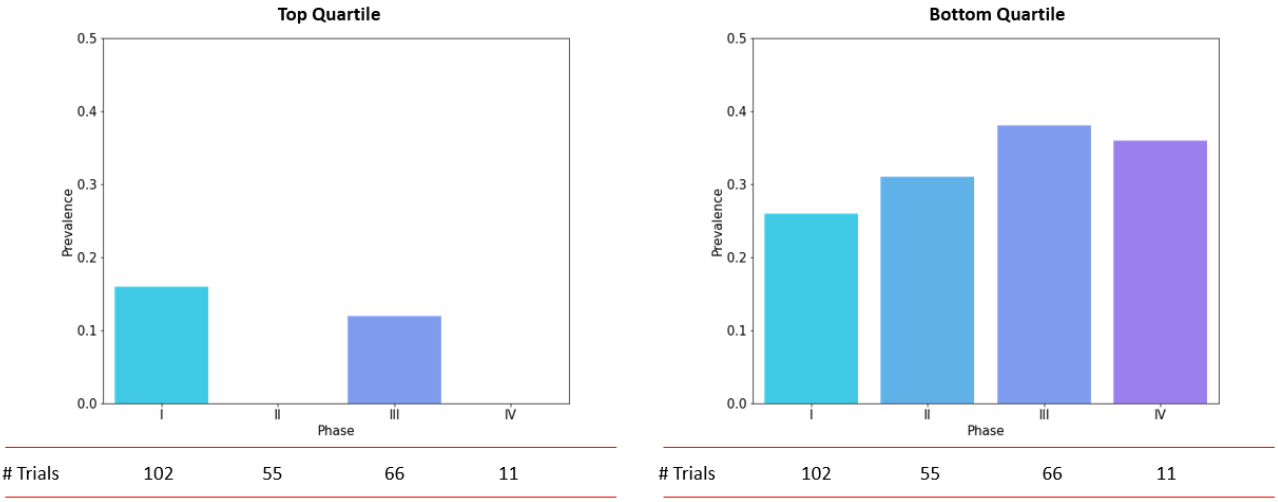


Figure 10 Prevalence of top quartile (left) and bottom quartile (right) trial performance by Phase

Finally, we can observe prevalence across trial phases in *Figure 10*. Once again, the low prevalence of top quartile trials is noticeable, alongside the considerably higher prevalence of bottom quartile trials. The average prevalence across phases is lower than across either therapeutic area or CRO. This may indicate that reducing these cycle times is less about the structure of the testing phase, and more about the requirements of each therapeutic area or partners involved. As initially noted in the boxplot analysis, the correlation between trial phase and close out slowdown is further observed. Bottom quartile trials tend to be more frequent with increasing trial phase, with no top quartile trials occurring in Phase II or IV.

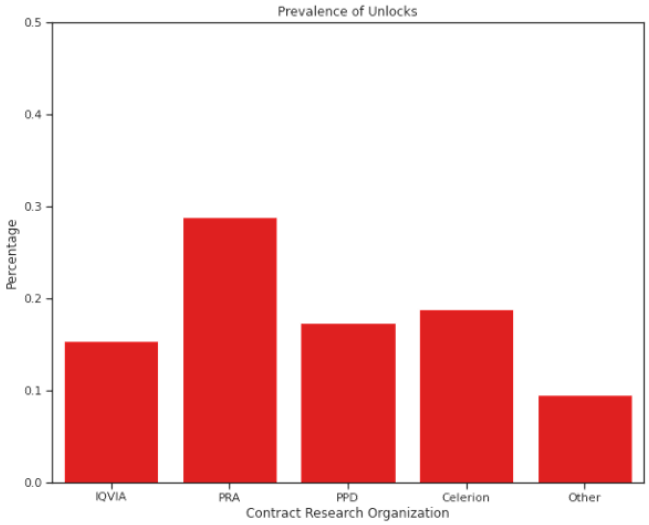
The aim of the plots above is to provide a deeper insight into what the causes for top or bottom quartile performance may be. The plots were also positively welcome by our counterparts at Takeda as these allowed for more targeted conversations focusing the trials with the lowest prevalence of top quartile performance (Phase I, Oncology, IQVIA, PPD, etc.).

Exploratory Data Analysis: Data Unlocks

As part of the mission to improve the quality of data generated during clinical trials, understanding drivers for data unlocks may provide a starting point. A database unlock typically occurs following from an initial database lock at the end of a clinical trial. While there may be a multitude of specific reasons justifying an unlock, these occur when a data review has been triggered – usually resulting from incorrect or missing values, need in the statistical

analysis. Unlocks can lead to a considerable slowdown in cycle times as these restrict and reset any advanced data analysis being carried out.

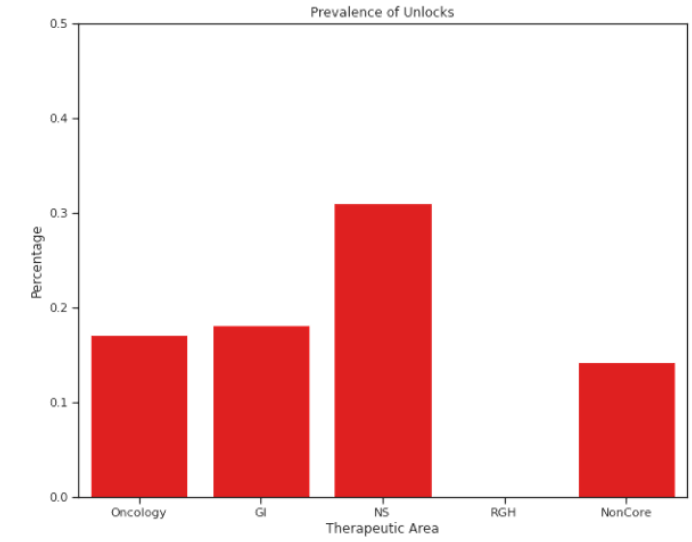
In generating database unlock insights, we leveraged a dataset recording information on trial unlocks from 2019 to 2021. The starting date of this dataset is due to Takeda's internal decision to begin recording these unlocks in 2019.



# Trials	26	38	23	16	21

Figure 11 Data Unlock prevalence broken down by CRO

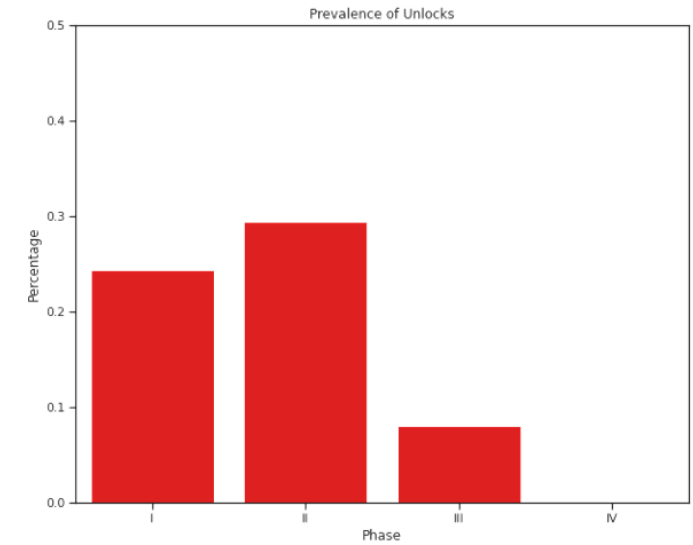
Figure 11 provides the unlock prevalence breakdown by CRO. Prevalence here is defined as the ratio of instances of unlocks observed for a given category -CRO, in this case- to the total number of trials carried out in this category. As per *Figure 11* it appears that PRA has the highest frequency (30%) of unlocks. This spotlights PRA as a partner who, despite posting shorter cycle time, requires frequent unlocks to reanalyze the data – thereby further slowing down the pipeline. Celerion, in a similar manner, shows the second-highest frequencies (20%), despite initially presenting shorter cycle times. Notably, the *Other* category of CROs – typically university research centers, or smaller scale research facilities- have the lowest frequency of unlocks.



# Trials	41	33	29	7	17

Figure 12 Data Unlock prevalence by Therapeutic Area

Figure 12 provides a similar plot, with unlocks subset by Therapeutic Area. The neuroscience division (NS), the best performer according the boxplots in *Figure 6,* demonstrates the highest frequency of unlocks. Conversely, the rare diseases unit (RGH) does not appear to have had any data unlocks between 2019 and 2021, although only seven trials were carried out in this time frame.



# Trials	45	34	37	8

Figure 13 Data Unlock prevalence by Phase

In plotting the prevalence across phase, in *Figure 13,* we notice an opposite trend to that observed *Figure 7*. Despite cycle times generally increasing with increasing phase, the frequency of unlocks tends to decrease. Phase I and II both displaying frequencies above 20%, while phases III and IV, despite involving a larger patient population, have frequencies of 8% and 0%, respectively.

The results above suggest that deeper dives are required to investigate drivers for data unlocks. Whether through descriptive or predictive approaches, additional data is to be explored to fully gauge true signals in the operations. In our conversations with Takeda, we have outlined a goal to explore the use of machine learning to predict whether, for a given trial, an unlock may occur. The sections below will focus on our findings deriving from our machine learning efforts, alongside any key takeaways from our predictive analyses.

Machine Learning Insights

We started the machine learning analysis from our regression part of predicting Close Out time. To do that, we conducted a validation procedure where we trained our model on the training set and checked its performance on the validation set to select optimal hyperparameters for each model. In the end, we retrained all models with best hyperparameters on concatenated train and validation sets and evaluated performance on our test set just once. In-sample and out-of-sample metrics are reported below.

Table 1 Summary statistics for Close-Out time predictions

Model Name	Train MSE	Test MSE	Train MAE	Test MAE	Train R^2	Test R^2
Linear Regression	1094	2423	24.80	37.40	0.334	-0.148
Lasso Linear Regression	1428	2702	28.83	40.07	0.132	-0.281
Ridge Linear Regression	1190	2350	25.95	36.23	0.276	-0.113
Decision Trees	1395	2748	27.92	40.63	0.151	-0.302
Gradient Boosting	359	2399	13.56	38.77	0.781	-0.137
Poisson Regression	1645	2674	32.38	41.80	0.001	-0.267

We can observe that our Out-of-sample R^2 values are less than 0 across the board, reflecting that even predicting an average time would work better than any of our models. Acknowledging this, we decided to move from trying to predict the future to understanding what features affecting the Close Out time are statistically significant.

For this part we conducted interpretable machine learning in depth by gathering information on statistically significant features, parameters, and confidence intervals. To do this, we found a best L1 hyperparameter (alpha) from a validation stage and applied it to find all coefficients and parameters from Lasso Linear Regression. Our summary results are presented in *Figure 13*.

OLS Regression Results

Dep. Variable:	CloseOut_new	R-squared:	0.169
Model:	OLS	Adj. R-squared:	-0.013
Method:	Least Squares	F-statistic:	0.9297
Date:	Fri, 13 Aug 2021	Prob (F-statistic):	0.589
Time:	16:24:15	Log-Likelihood:	-1041.0
No. Observations:	207	AIC:	2158.
Df Residuals:	169	BIC:	2285.
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PlannedEnrollment	0.0442	0.026	1.706	0.090	-0.007	0.095
ActualEnrollment	-0.0076	0.027	-0.285	0.776	-0.060	0.045
ActualRandomizedDosed	-0.0106	0.032	-0.326	0.744	-0.075	0.053
ActualEarlyTerminated	0.0128	0.031	0.411	0.681	-0.049	0.074
PercEnrollmentCompleted	0	0.013	0	1.000	-0.025	0.025
PlannedSites	0.0095	0.071	0.135	0.893	-0.130	0.149
ActiveSites	-0.0330	0.269	-0.122	0.903	-0.564	0.498
numValsLSO	0	1.891	0	1.000	-3.733	3.733
numValsDBL	0	1.891	0	1.000	-3.733	3.733
PrimaryStudy_to_Synopsis	-0.0510	0.019	-2.724	0.007	-0.088	-0.014
Synopsis_to_Protocol	-0.0057	0.020	-0.283	0.778	-0.046	0.034
FSS_to_FSI	0	0.239	0	1.000	-0.472	0.472
LSI_to_LSO	-0.0028	0.014	-0.199	0.843	-0.030	0.025
AE	0	14.479	0	1.000	-28.583	28.583
CM	0	46.792	0	1.000	-92.373	92.373
DS	0	14.479	0	1.000	-28.583	28.583
DV	0	24.458	0	1.000	-48.283	48.283
EG	0	26.781	0	1.000	-52.867	52.867
EX	0	14.479	0	1.000	-28.583	28.583
FA	0	18.659	0	1.000	-36.835	36.835
LB	0	14.479	0	1.000	-28.583	28.583
MH	0	14.479	0	1.000	-28.583	28.583
PR	0	22.705	0	1.000	-44.822	44.822
SS	0	19.735	0	1.000	-38.958	38.958
VS	0	31.071	0	1.000	-61.338	61.338
Lock_duration	1.5777	0.609	2.593	0.010	0.376	2.779
Expected_CloseOut	0.2863	0.070	4.069	0.000	0.147	0.425
Ontime	0	8.918	0	1.000	-17.605	17.605
Unlocks	0	18.160	0	1.000	-35.849	35.849
COVID	0	19.644	0	1.000	-38.779	38.779
0	0	15.517	0	1.000	-30.633	30.633
I	0	6.979	0	1.000	-13.777	13.777
II	0	7.698	0	1.000	-15.196	15.196
III	0	8.393	0	1.000	-16.568	16.568
IV	0	13.876	0	1.000	-27.393	27.393
GI	0	6.887	0	1.000	-13.596	13.596
NS	0	7.320	0	1.000	-14.450	14.450
NonCore	0	6.554	0	1.000	-12.939	12.939
Oncology	0	6.694	0	1.000	-13.215	13.215
RGH	0	10.041	0	1.000	-19.822	19.822
PercActualCompetedTrial	0.0004	0.001	0.344	0.732	-0.002	0.003
Celerion	0	15.020	0	1.000	-29.650	29.650
IQVIA	0	7.153	0	1.000	-14.121	14.121
Other	0	6.500	0	1.000	-12.831	12.831
PPD	0	6.891	0	1.000	-13.604	13.604
PRA	0	7.474	0	1.000	-14.754	14.754

Omnibus:	30.268	Durbin-Watson:	1.810
Prob(Omnibus):	0.000	Jarque-Bera (JB):	41.064
Skew:	0.904	Prob(JB):	1.21e-09
Kurtosis:	4.222	Cond. No.	1.16e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.17e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 13 Summary Statistics from the best Lasso Linear Regression

We observed that our Adjusted R^2 = 0.169 which indicates that with current set of available parameters for clinical trials we cannot explain a major part of variance. However, we identified two statistically significant drivers of close out time:

- **Lock Duration**: the duration of time between unlocking data and relocking data took. This parameter is significant at a 95% confidence interval which from 0.376 to 2.779. The average value of the parameter is 1.57 - more than 1. This means that every additional day of relocked data, on average, adds 1.57 days to the final close out time.

- **Expected Close Out:** the expected duration of this segment calculated for each clinical trial prior to a trial starting. We found this parameter to also be significant at a 95% confidence interval, taking values between 0.147 to 0.425 with an average value of 0.286. It means that every additional day in expected close out, on average, adds to 0.286 to the actual close out time.

The information above is valuable to Takeda for two main reasons:

1. Takeda now has a quantified understanding of how much data relocks affect the final performance during close out. On top of that, Takeda is now aware that one additional day of relock adds more than one day in actual close out time. This information can be used to quantify impact upon minimization of unlocks in the future.

2. Takeda knows that current initial expected close out time doesn't translate perfectly to an actual close out time, leaving room for improvement in terms of predictive capabilities.

Since accurately predicting the exact time of Close Out is a complex problem, we decided to switch our attention to understanding how well Takeda performs at meeting its internal deadlines. We decided to predict whether, for each clinical trial, the actual Database Lock date occurred on or before the expected date.

Below are our summary results for the classification analysis in *Table 2*. We achieved a strong Out-of-sample ROC AUC of 0.821 reflecting a good predictive power in terms of separating between on-time and not on-time delivery of clinical trials.

Table 2 Summary statistics for On-time close-out delivery

Model Name	Train ROC AUC	Test ROC AUC	Train F1 Score	Test F1 Score	Train Accuracy	Test Accuracy
Logistic Regression	0.881	0.714	0.864	0.410	0.824	0.603
Decision Trees	0.500	0.500	0.797	0.343	0.663	0.207
Random Forest	0.969	0.732	0.947	0.435	0.929	0.552
Gradient Boosting	0.549	0.496	0.806	0.353	0.682	0.241
Categorical Boosting	1.000	0.821	1.000	0.478	1.000	0.586

We would also want to mention that even an interpretable model shows Out-of-sample ROC AUC of 0.714 which is much better than a random assignment of classes. Since we care about interpretability of our models, we wanted to look at feature importance of our best interpretable model – summary statistics can be found on *Figure 14*.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.130
Dependent Variable:	Ontime	AIC:	309.7114
Date:	2021-08-11 21:05	BIC:	355.7478
No. Observations:	255	Log-Likelihood:	-141.86
Df Model:	12	LL-Null:	-163.00
Df Residuals:	242	LLR p-value:	2.9884e-05
Converged:	1.0000	Scale:	1.0000
No. Iterations:	30.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
PlannedEnrollment	-0.0036	0.0015	-2.3629	0.0181	-0.0065	-0.0006
ActualEnrollment	0.0010	0.0017	0.5876	0.5568	-0.0023	0.0044
ActualRandomizedDosed	0.0072	0.0041	1.7627	0.0779	-0.0008	0.0152
ActualEarlyTerminated	-0.0068	0.0052	-1.2952	0.1952	-0.0171	0.0035
PercEnrollmentCompleted	-0.0019	0.0021	-0.8942	0.3712	-0.0060	0.0022
PlannedSites	0.0148	0.0057	2.5856	0.0097	0.0036	0.0260
ActiveSites	0.0462	0.0442	1.0464	0.2954	-0.0403	0.1328
Synopsis_to_Protocol	0.0006	0.0009	0.7399	0.4593	-0.0011	0.0024
FSS_to_FSI	0.0115	0.0091	1.2683	0.2047	-0.0063	0.0293
LSI_to_LSO	-0.0000	0.0005	-0.0166	0.9867	-0.0011	0.0010
AE	0.0000	nan	nan	nan	nan	nan
CM	0.0000	nan	nan	nan	nan	nan
DS	0.0000	nan	nan	nan	nan	nan
DV	0.0000	nan	nan	nan	nan	nan
EG	0.0000	nan	nan	nan	nan	nan
EX	0.0000	nan	nan	nan	nan	nan
FA	0.0000	nan	nan	nan	nan	nan
LB	0.0000	nan	nan	nan	nan	nan
MH	0.0000	nan	nan	nan	nan	nan
PR	0.0000	nan	nan	nan	nan	nan
SS	0.0000	nan	nan	nan	nan	nan
VS	0.0000	nan	nan	nan	nan	nan
Lock_duration	-0.0323	0.0203	-1.5876	0.1124	-0.0721	0.0076
Expected_CloseOut	0.0115	0.0034	3.3337	0.0009	0.0047	0.0182
PercActualCompetedTrial	-0.0019	0.0018	-1.0995	0.2716	-0.0054	0.0015
Celerion	0.0000	nan	nan	nan	nan	nan
IQVIA	0.0000	nan	nan	nan	nan	nan
Other	0.0000	nan	nan	nan	nan	nan
PPD	0.0000	nan	nan	nan	nan	nan
PRA	0.0000	nan	nan	nan	nan	nan
I	0.0000	nan	nan	nan	nan	nan
II	0.0000	nan	nan	nan	nan	nan
III	0.0000	nan	nan	nan	nan	nan
IV	0.0000	nan	nan	nan	nan	nan
GI	0.0000	nan	nan	nan	nan	nan
NS	0.0000	nan	nan	nan	nan	nan
NonCore	0.0000	nan	nan	nan	nan	nan
Oncology	0.0000	nan	nan	nan	nan	nan
RGH	0.0000	nan	nan	nan	nan	nan

Figure 14 Summary Statistics from the best Regularized Logistic Regression

We have found three statistically significant drivers of variability in delays (since in our classification 1 means on time delivery and 0 means off-time delivery):

1. **Planned Enrollment** – statistically significant decrease in the probability of being on time with increase in planned enrollment. This is logical since planned enrollment can be one of the proxies for the complexity of a trial.

2. **Planned Sites** – statistically significant increase in the probability of being on time with increase in planned number of sites. This can happen if certain sites are better clinical and data operators, and their inclusion may positively affect probability of being on time with close out delivery.

3. **Expected close out** – statistically significant increase in the probability of being on time with an increase in expected close out. This makes sense since allocating more time for close out gives sufficient time to properly review data, as opposed to unlocking it after the trial.

This information points Takeda to a direction where potential improvements can occur. Specifically, conducting a review of the operational performance of historical sites used by Takeda may help identifying which sites are best to partner with – if this choice can be made by Takeda.

Business Impact

The primary result that captured our attention was the coefficient assigned to the value of each unlock day. With a high degree of statistical significance, our machine learning models attributed a value of 1.57 extra days to close out for every day spent relocking an unlocked database. Although the significance of data unlocks was known and stated by Takeda stakeholders during our interactions, the impact of these on the entire trial pipeline was never truly quantified. Our work thus indicates that reducing data unlocks is a key matter if time to completion in trials is to be minimized. Being now aware of a 1.57-day slowdown resulting from a single day of relock, we can calculate the impact of reducing all relock times to zero – i.e., the data collection processes were correctly carried out in their entirety. We find that if no unlocks had occurred in the 2019-2021 window, each trial would have saved an average of 6.05 days out of 109 days required for a full close out. Reducing relock instances to zero thus speeds up our identified segment's cycle time by 5.5%.

Our conversations with Takeda highlighted that a speed up in clinical trials leads to an increase in the project's net present value (NPV) due to future cash flows resulting from the potential sales of the drug candidate. Internal Takeda resources estimate this NPV uplift to equal $ 30 million for every 30 days saved. Wong, Siah, and Lo [1] and the Biotechnology Innovation Organization (BIO) [2] can provide us with a success probability of a drug candidate broken down by both phase and therapeutic area. Thus, since we know the therapeutic area, phase, and the average number of days saved per trial (~1.6 days per 1 day of relock), we can calculate the expectation NPV lift if these relock days are minimized to zero. Since the relock coefficient stands at a 95% confidence interval, we can repeat this calculation for both upper

and lower value bounds thereby providing Takeda with a range of estimated expected yearly NPV uplifts, shown below.

Table 3 Breakdown of NPV calculations

	Worst Case	Conservative Case (99.9% confidence interval)	Best Case
Coefficient	0.376	1.577	2.779
Expected Yearly NPV uplift	$ 9 million	$ 39 million	$ 68 million

For the first time, Takeda now has a more refined understanding of how it can invest its time resources during a project. It is now clear that assigning more time to the close out phase can prevent data unlocks, which result in a truly meaningful impact on the drug candidate's trial economics.

5. Hurdles

Data Acquisition

Throughout the first leg of the capstone engagement (February-May) the main hurdle we faced was ensuring correct data and technology access. This included access to the correct software, as well as the correct operational data tables. In addition to the data access, the project scope was more general than it currently is, inherently leading to initial, higher-level discussions, as opposed to more detailed "hands-on" presentations.

To mitigate and overcome the hurdles above, we found that being clear and open about our needs was helpful. Being honest about the potential analytics tools and methods we could deploy within the organization helped us both shape any meetings to focus on our urgent needs as well as what we could achieve once these needs were met. In this regard, Takeda has been highly supportive in ensuring a streamlined experience. Once our technical requirements were clarified, the discussions for a slimmed project scope became more natural, and we were able to efficiently define the current project scope.

Finally, as we were waiting on our access requests to be fulfilled, we spent time between ourselves, with Takeda, and with Prof. Graves, whiteboarding potential solutions. Taking time to observe and digest the bigger internal vision Takeda had outlined helped us in outlining a plan of action we could act upon once we had been greenlighted by the internal IT teams.

The outlined progress in the previous sections, is all part of the whiteboarding and discussions we held in late May and early June, a testament that the mitigations our team and Takeda implemented have worked.

Close Out Analysis

In this analysis, we had two main hurdles:

1. **Creation of a relevant dataset.** One of the main problems of the analysis was that the appropriate dataset had not been created earlier, and information about where what data was located needed to be learned from different people. To make the dataset, we needed to conduct a series of meetings with members of the GDO and the Takeda Data Science Institute, who explained to us how the pipeline for receiving data from different CROs works and how we can access it. Since the relevant dataset did not originally exist within the Takeda system, we had to create each clinical study parameter manually from different tables. After making the relevant parameters, we merged the parameters by the unique ID of the clinical trials.

2. **Finding the signal in the data.** With an initial set of parameters for each clinical trial, we found that Out-of-Sample quality results were meagre and, in most cases, negative. To solve this problem, we completed the following:

    a. ***Creation of additional parameters.*** We had a series of meetings with Takeda. We discussed what is going on between the LSO and DBL and what different parameters can be collected after receiving the final data from the CRO. Thus, we could find data that were not available with the initial level of access for our analysis.

    b. ***Choosing a different Machine Learning Design.*** Since initially, we formulated the problem as a regression with Close Out time from zero to infinity, we decided to try to change the problem set (1) to multiclassification, where we tried to predict the levels of the dependent variable from the lowest value to the largest or (2) binary classification when we tried to predict whether Close Out high or low. The results are presented in detail in the Results section.

Data Unlocks Analysis

In this analysis, we also had three main hurdles:

1. **Identify a specific problem.** Over several months, we have been trying to identify with the Takeda team a particular problem that we could solve related to Patient Data. The initial setting of the problem was too general, which stopped us from progressing on the issue. As a result, we held several meetings with the Patient Data team, after which we were able to crystallize the problem formulation.

2. **Get relevant data.** There is no pertinent data in the shared datahub indicating Unlocks. To solve this problem, we presented what were at the time our results and problems for the joint Operation Data, Patient Data and Global Development Office teams, which made it possible to learn about obtaining data effectively. This allowed to access the data only one day after meeting with the teams.

3. **Find a signal in the data.** Initial results also showed low Out-of-sample quality metrics. To solve this problem, we used a wide range of potential models ranging from interpretable to black box. In addition, we tried to change the problem statement from predicting the number of Data Unlocks to a binary target showing whether the number of Unlocks was greater than or equal to 0.

6. Deliverables and Transition Plan

Deliverables

Resulting from our work, we have outlined and discussed -along with Takeda- a set of deliverables which, we find, will be of most benefit to Takeda. Each of the deliverables outlined below is representative of the three stages we worked on throughout the project: data engineering, descriptive and predictive analytics, and subsequent recommendations.

- *Data Access Queries and Wrangling (SQL and Python Notebooks)*
    - Our data engineering efforts required most of our time resources throughout the project as we had to constantly vet the sources of truth as well as the quality of the data analyzed. Developing an adequate data table led us to generate numerous SQL queries alongside relevant Python data wrangling Jupyter Notebooks. Following from the correct formatting and documentation of our work, we plan on sharing the developed queries with Takeda's team of data analysts to ensure that there is a clear understanding as to how our data was generated.

- *Machine Learning Models (Python Notebooks)*
    - As per the previous sections in this report, numerous machine learning frameworks were applied through a series of models. The drafting and implementation of the models -alongside the relevant results interpretation- was all completed on Jupyter notebooks, with code written in Python. To provide an understanding behind the implementation of the models (including data processing and model selection) we aim to share these annotated notebooks with the relevant stakeholders.

- *Recommendations and Insights (PowerPoint presentation)*
    - Upon the conversion to our full-time roles, we maintained a rolling cadence of weekly touchpoints (15-30 minutes) and biweekly presentations (45-60 minutes). For the most part, these meetings included slides used as means of visualizing and interpreting our results. Each insight was provided with the required explanation, delivered either verbally or through text on the slides themselves. To ensure that Takeda retains all generated insights and explanations, we aim to share formatted slides that walk relevant and interested stakeholders through our work.

Transition Plan

Of essential importance -if our work is to generate practical value- is the need for an adequate transition plan and knowledge transfer process. Following from strong excitement by Takeda's stakeholders, we have organized knowledge transfer meetings – to be held during the final week of our capstone engagement (16-20th August). Through these sessions we will provide a deep-dive explanation of the structure of our data generation and analysis pipeline. We will also use this opportunity to summarize our recommendations and provide potential next steps, to be undertaken either internally by Takeda or through a future capstone project. Although we have regularly updated all stakeholders with a reasonable frequency throughout our engagement, these knowledge-transfer sessions will be more practical and hands-on.

Given the number and seniority of stakeholders our work has engaged over the course of the past months, our meetings will take place with both senior leaders and as well as junior colleagues. Our objective is for these meetings to ensure that there is an across-the-board understanding of our work such that there is minimal friction in continuing our work and that any project vision can be refined.

7. Concluding Thoughts

Through these months of engagement, numerous meetings, brainstorming sessions, and data efforts, our learnings have certainly exceeded our expectations. Having started with a minimal understanding of the pharmaceutical industry and clinical trials, we now conclude with a considerably more refined view and appreciation for the work conducted in this field. We are highly excited by the path Takeda has envisioned to follow and look forward to learning more about their future work. While diving deep into our technical and professional lessons would require an additional report then length of this very one, we'd like to conclude our report with summarizing points encapsulating our experience.

- Any complex problem can truly be simplified into smaller, solvable tasks
- Data quality is arguably the most important factor in deploying impactful machine learning
- Vision and ambition are essential to generate value in large organizations

References

1. Wong, C. H., Siah, K. W., & Lo, A. W. (2018). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*(2), 273–286. http://doi.org/10.1093/biostatistics/kxx069
2. Clinical Development Success Rates 2006-2015 (2016) Biomedtracker. https://www.bio.org/sites/default/files/legacy/bioorg/docs/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf