Predictive Modeling for Clinical Trial Recruitment
A Mosaic Data Science Case Study

Clinical Trial Recruitment | An Opportunity for Data Science

The only way to get new medicines to market is to run them through a clinical trial. After a potentially long and expensive drug development period and in the face of rising clinical trial costs, it is no surprise that pharmaceutical firms invest substantially in designing the perfect trial. Even so, close to 80% of trials face completion delays. According to the website Pharmafile, the financial impact of clinical trial delays can be substantial: losses of $0.6M–$8M per day in subsequent sales can be attributed to these delays. While there are various causes of clinical trial delays, Intralinks found that delayed patient recruitment & enrollment caused study delays in 41% of trial sites, making it the second-leading cause of such delays.

For many pharmaceutical firms, trial recruitment forecasting plays a role in trial recruitment planning. However, these forecasts may be generated with relatively simplistic approaches based on only a small subset of available internal & external data. Their poor performance decreases trust in them among trial planners, who, in the absence of dependable forecasts, often succumb to the natural tendency to set relatively optimistic trial plans. Trial completion delays and corresponding financial losses and damaged relationships ensue. In the rare cases where trial recruitment plans are set too conservatively, resources and budget are over-allocated to the trial. Dependable recruitment forecasts can enable more realistic recruitment expectations, leading to improvements in decisions related to clinical trials, such as the selection of a baseline trial recruitment plan, how many and which sites and investigators to select for a trial, and when and how to intervene to improve recruitment during a trial.

After struggling with expensive clinical trial completion delays and a lack of trust in recruitment forecasts from an off-the-shelf tool that inhibited a more quantitative approach to trial planning, one of the world's largest pharmaceutical companies sensed an opportunity to leverage data science. When the company was not sure how to start leveraging additional data and more sophisticated data science techniques, they reached out to Mosaic, a leader in AI consulting, to assist with initial efforts.

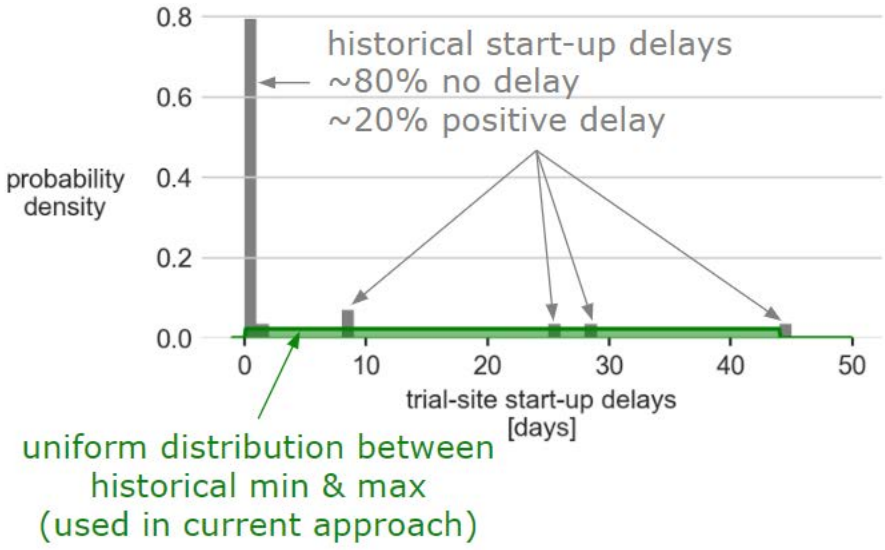
Moving Quickly

At the beginning of the project, Mosaic collaborated with stakeholders to determine success criteria, assess internal & external data sources, and investigate potential modeling and forecasting solutions. Stakeholders identified three main uses for improved clinical trial recruitment forecasts and a wide range of possible data sources. A review of relevant literature revealed a range of possible modeling approaches. Rather than embarking on a prolonged data engineering effort, extensive research into modeling approaches, and building a solution to serve multiple use-cases, Mosaic and the stakeholders chose to shorten time-to-value and increase early learnings by starting with the most promising and readily available internal & external data and a relatively straightforward predictive modeling approach that adjusted for the most glaring deficiencies in the current off-the-shelf forecasting approach, as well as focusing on just one use-case for the forecasts. After less than 6 months of part-time work by a small team, Mosaic and the company demonstrated enough promise to justify additional investment in deployment of the new approach in a prototype dashboard.

Copyright © Mosaic Data Science. All rights reserved.

Exploratory Data Analysis | First Step to Predictive Model Development

Mosaic follows the CRISP-DM process for most analytics project. An exploratory data analysis (EDA) is critical to understanding the data, evaluating potential new sources, and getting data ready for predictive modeling. After spending time with stakeholders, Mosaic's data scientists looked for anomalies, identified trends, visualized the data, and began feature engineering to get the data ready for a predictive model. Some EDA results were no surprise, such as the fact that trials rated as more complex were more likely to experience startup delays. Others were unexpected, such a demonstration via hierarchical regression analysis that variations in recruitment performance depended more on differences in trials than in sites or investigators. The off-the-shelf forecasting tool only used site data when building forecasts, suggesting an opportunity for improvement. Another such opportunity revealed itself during EDA when the distribution of startup delays was found to differ substantially from the distribution assumed by default in the off-the-shelf tool, as shown in the figure below.



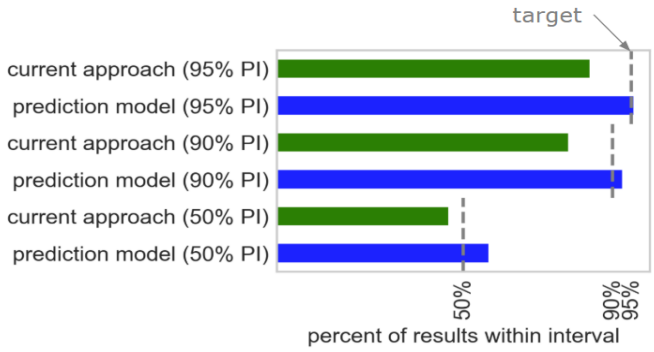
Improving Accuracy of Trial Recruitment Forecasts

After reviewing relevant literature and the approach used by the off-the-shelf tool, Mosaic selected Generalized Linear Model (GLM) predictive models of recruitment forecast parameters as the core of the forecasting approach. GLMs are a powerful, flexible, and interpretable approach that predict a full probability *distribution* of the target or outcome variable conditioned on the input feature values. Predicting *distributions* was essential in this context because multiple predicted forecast parameter distributions are combined to produce forecasted recruitment distributions. These could be used to produce prediction intervals or predicted probabilities of various outcomes. Producing forecasted distributions, though more challenging than producing point forecasts, was essential to the company's trial decision-making processes, which consider the chances of various outcomes.

In order to evaluate the performance of the new forecasts relative to the current approach, Mosaic reverse-engineered the off-the-shelf solution and implemented an approximation of it. Forecasting approaches were compared to the actual recruitment at sites and in trials in the test data set. The new forecasting approach produced errors in the count of patients screened at trial-sites that were 33%

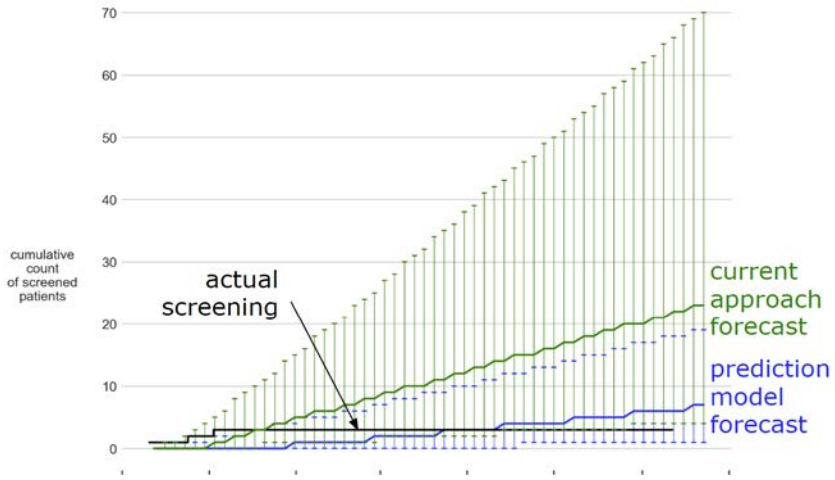
Copyright © Mosaic Data Science. All rights reserved.

lower than the current approach. Furthermore, the approach demonstrated superior prediction interval calibration, as shown in the figure below.



Mentorship & Collaboration While Getting Forecasts to Decision Makers

In an effort to integrate the new forecasts into decision-making processes and learn from user feedback, a prototype dashboard is being developed. A draft interface provided by subject-matter experts on the team will be adjusted as needed based on analysis and modeling results. The dashboard itself will be produced by a dashboarding expert at the pharmaceutical company, with the enabling data and models provided by Mosaic and a growing internal data science team at the company. As the project has progressed, Mosaic has begun mentoring new data scientists on that team and working collaboratively with them on additional data set integration and model type explorations, as well as model deployment. This mentorship and collaboration will ensure that the new internal team is ready to take ownership of the data and modeling behind the prototype dashboard after it is deployed.



The image above visualizes the efficiency of Mosaic's forecast output.

For More Information

Want to learn more? Please contact info@mosaicdatascience.com

Copyright © Mosaic Data Science. All rights reserved.