# PREDICTIVE MODELING *for* CLINICAL TRIAL RECRUITMENT

MOSAIC DATA SCIENCE | CASE STUDY

## MACHINE LEARNING CASE STUDY

Predictive Modeling for Clinical Trial Recruitment

*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.*

## 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.

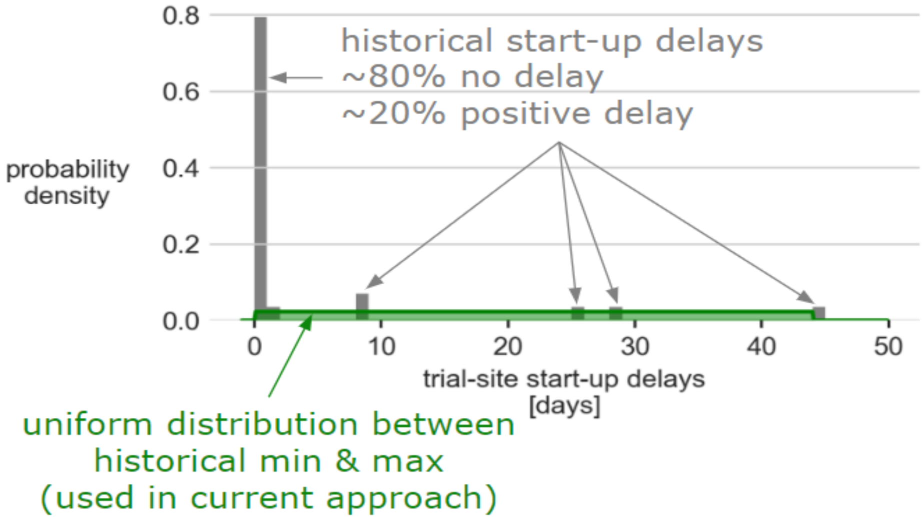
After struggling with expensive clinical trial completion delays and a lack of trust in recruitment forecasts from an off-the-shelf tool that inhibited a more quantitative approach to trial planning, one of the world's largest pharmaceutical companies sensed an opportunity to leverage data science. When 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.

## EXPLORATORY DATA ANALYSIS | FIRST STEP TO PREDICTIVE MODEL DEVELOPMENT

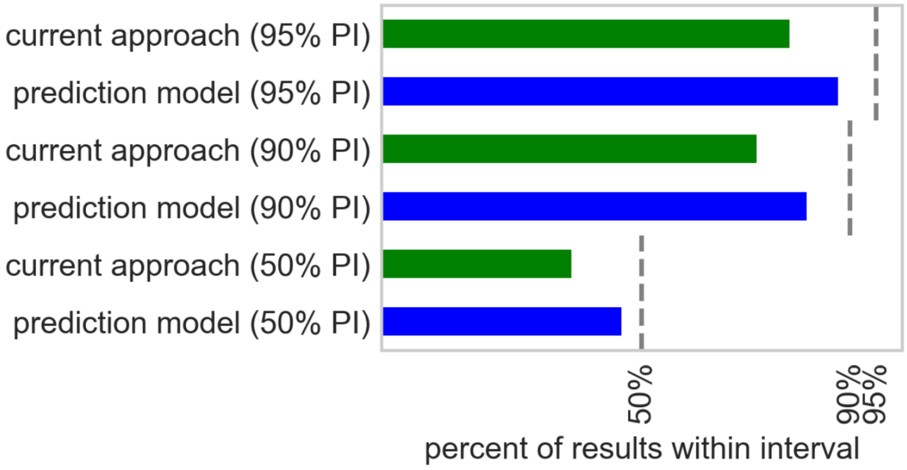
Mosaic follows the CRISP-DM process for most analytics projects. An exploratory data analysis (EDA) is critical to understanding the data, evaluating potential new sources, and getting data ready for predictive modeling. After spending time with stakeholders, Mosaic's data scientists looked for anomalies, 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 on differences 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 class of models 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, such as the probability of achieving a trial recruitment goal by a certain date. Producing forecasted distributions, though more challenging than producing point forecasts, was essential to the client's trial decision-making processes, which consider the chances of various outcomes. The client data science team used an enterprise automatic machine learning platform to find models that produced more accurate point predictions of site delays or rates than GLMs, but GLMs were selected for deployment because the other models were not better enough at producing forecasted distributions to be worth their lower interpretability.

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 20% lower than those of the current approach. Furthermore, the new forecasting approach demonstrated superior prediction interval calibration, as shown in the figure below.



## SITE SELECTION AND OPTIMIZATION

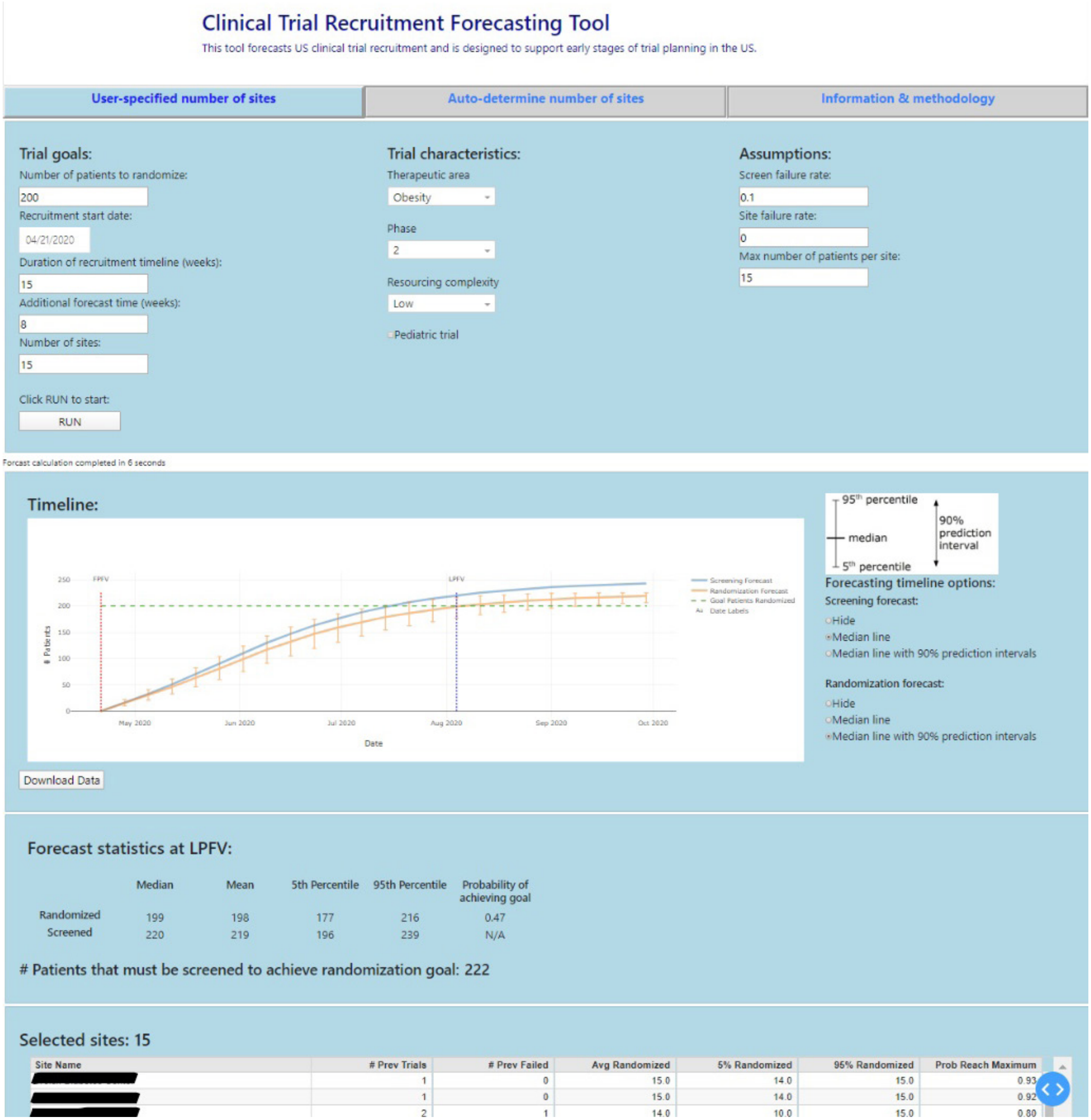
While more accurate and better calibrated forecasts of the number of patients that a site will recruit in an upcoming trial are useful in their own right for site monitoring and other activities, important trial planning decisions regarding the planned duration of the trial or the goal number of patients to recruit require trial-level forecasts, and each trial involves many sites. Mosaic developed trial-level forecasts while simultaneously proposing sites for the trial, thereby also accelerating another trial planning activity: the selection of sites for the trial. Proposing sites for a trial is a prescriptive analytics optimization problem that Mosaic solved by building upon the site-level forecasts developed earlier. In particular, Mosaic worked with the client to specify two variations of the site selection problem. Both of these approaches depend critically on the predicted distributions of the number of patients recruited at each site over time produced by the site-level forecasting approach.
The first variation of the problem allowed the user to specify the number of sites, how long the recruitment period would last, and how many patients need to be recruited by all the sites in the trial. The optimization approach then  finds the set of sites that maximize the probability of achieving the goal number of patients for the trial by the end of the recruitment period. For example, a trial manager could input that they need to recruit 200 patients in 15 weeks with 20 sites, and the tool would select the optimal 20 sites and return the probability that those sites would reach that goal. If that probability was too low, the trial manager would know they might need to re-negotiate some aspects of the trial specifications. For example, they might request a longer recruitment period or the ability to use more sites.

The second site selection problem variation was more complex because it also finds an appropriate number of sites for the user. Instead of fixing the number of sites, the user specifies a probability of achieving the goal and then the optimization approach determines how many sites are needed to achieve the goal with at least that user-specified probability. For example, a trial manager could specify that they want to be 90% confident that they will recruit 200 patients in 15 weeks. The tool would then determine how many and what sites should be selected such that the goal is forecasted to be met with at least probability 0.9. This tool provides a new capability to trial managers that is not available in the current off-the-shelf solution. The number of sites for a trial has  been selected manually using heuristics; this tool allows more data-driven decision making regarding the number of sites in a trial.

## DASHBOARD DEVELOPMENT

In an effort to integrate the new forecasts into decision-making processes and learn from user feedback, Mosaic shifted into developing a dashboard. The team worked with subject-matter experts to define a draft interface to be made into a prototype dashboard. This country-wide prototype will provide significant value to the company because only one person is currently licensed to use the off-the-shelf forecasting tool, creating an information bottleneck. An accessible dashboard was key to unlocking the value of the new models Mosaic developed and enabling data-driven decision making across the organization.

Initially, the goal was to visualize the forecasts in Tableau. However, after an initial exploration, it was determined that the forecast generation process was too complex for Tableau to support. Mosaic stepped up to develop an alternative solution using an open-source dashboarding tool called Dash. Dash is a framework for building dashboard web applications in Python. Within two months the team had the prototype dashboard up and running.  From there, the dashboard was iteratively refined based on user feedback.



## Clinical Trial Recruitment Forecasting Tool

This tool forecasts US clinical trial recruitment and is designed to support early stages of trial planning in the US.

User-specified number of sites
Auto-determine number of sites
Information & methodology

**Trial goals:**

Number of patients to randomize: 200

Recruitment start date: 04/21/2020

Duration of recruitment timeline (weeks): 15

Additional forecast time (weeks): 8

Number of sites: 15

Click RUN to start:

**Trial characteristics:**

Therapeutic area: Obesity

Phase: 2

Resourcing complexity: Low

□ Pediatric trial

**Assumptions:**

Screen failure rate: 0.1

Site failure rate: 0

Max number of patients per site: 15

Forcast calculation completed in 6 seconds

**Timeline:**

**Forecasting timeline options:**

Screening forecast:

○ Hide
● Median line
○ Median line with 90% prediction intervals

Randomization forecast:

○ Hide
○ Median line
● Median line with 90% prediction intervals

**Forecast statistics at LPFV:**

| | Median | Mean | 5th Percentile | 95th Percentile | Probability of achieving goal |
|---|---|---|---|---|---|
| Randomized | 199 | 198 | 177 | 216 | 0.47 |
| Screened | 220 | 219 | 196 | 239 | N/A |

**# Patients that must be screened to achieve randomization goal: 222**

**Selected sites: 15**

| Site Name | # Prev Trials | # Prev Failed | Avg Randomized | 5% Randomized | 95% Randomized | Prob Reach Maximum |
|---|---|---|---|---|---|---|
| [redacted] | 1 | 0 | 15.0 | 14.0 | 15.0 | 0.93 |
| [redacted] | 1 | 0 | 15.0 | 14.0 | 15.0 | 0.92 |
| [redacted] | 2 | 1 | 14.0 | 10.0 | 15.0 | 0.80 |

*A screenshot of the prototype dashboard.*

## MENTORSHIP & COLLABORATION WHILE GETTING FORECASTS TO DECISION MAKERS

All along the way, Mosaic teamed up with the client's internal data science team.. As the project progressed, Mosaic began 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. By the time the dashboard was complete, Novo Nordisk data scientists were contributing to model and dashboard development. This mentorship and collaboration will ensure that the new internal team is ready to take ownership of the data, models and prototype dashboard so they can provide continued value to the organization.

### FOR MORE INFORMATION

*Want to learn more?*

Please contact info@mosaicdatascience.com

mosaic
DATA SCIENCE

## ABOUT MOSAIC DATA SCIENCE

**We provide innovative machine learning, AI and analytics consulting across organizations.**

Mosaic is a leading data science consulting company focused on helping organizations build and deploy actionable analytics solutions. Our customers are as varied as the techniques we use — some just starting their first predictive analytics project; others with deep in-house machine learning expertise.

## HOW WE WORK WITH YOU

We work in a highly collaborative partnership with our customers to ensure you get only the best results to consistently drive business value.

### MACHINE LEARNING

*We design and deploy predictive algorithms to solve the most challenging problems facing businesses today*

### ARTIFICIAL INTELLIGENCE

*We bring a wealth of knowledge on how to tune AI models to deliver the maximum business value*

### BUSINESS ANALYTICS

*We leverage technical expertise and experience across a swath of industries, bringing fresh approaches to challenging problems*

mosaic
DATA SCIENCE

mosaicdatascience.com ✉ info@mosaicdatascience.com ☎ (866) 202-8600