Jess Weng, Mahirah Syakhroza, Grace Lee

Regeneron Team Bravo Final Report

The Problem

Regeneron Pharmaceuticals, a leading biotechnology company, develops life-transforming medicines for various illnesses. As advances in molecular biology and medical imaging drive rapid change in healthcare, clinical trials have evolved with new formats, endpoints, biomarkers, and data sources—whilst also becoming longer and more complex. To address the growing need for efficient trial evaluation, this project aims to develop an interactive visualization tool using ClinicalTrials.gov data to standardize, benchmark and analyze endpoint complexity across trials by reviewing data, defining complexity metrics, and implementing comparative analysis.

The Data & Methodology

This project uses the live AACT database from ClinicalTrials.gov (connected directly to Preset to ensure real-time updates and long-term scalability). In the initial dashboard, we focused on industry-sponsored interventional trials completed within the past 10 years, individually filtering the data for each chart to enhance relevance and model performance. In our final dashboard, we included all data and left the filtering options up to the user so they can freely explore all connections.

To ensure consistency, we leveraged AACT's version of a MeSH hierarchy to map multiple labels into standardized categories, preserving critical information to allow for more accurate benchmarking in our final visualizations. However, the raw database still contained many null values, outliers, inconsistent labels, and free-text fields. We therefore standardized the data with specific SQL queries: calculating trial duration by subtracting start and completion dates, removing incomplete entries and using keyword/punctuation matching statements to reorganize and group fields like inclusion criteria, outcomes, locations, conditions, and sponsor names.

The Results

We created multiple charts to illustrate the dynamics between termination rates by condition, termination rates by inclusion criteria, studies by source, countries by phase, countries by active clinical trials, range per phase, clinical trials by year, sponsor relationship to phase, most common primary outcome measures and many others.

Users on the enterprise Preset.io plan will be able to import the dashboard data and filter by study phase, term, country, and other key attributes—enhancing the usability and depth of insights provided by the dashboard. The entire project is available on Github with the full set of SQL queries, user stories/presentation slides and demo video.