Building an ML model to identify undiagnosed patients with rare disease

Life sciences company partners with Clarivate Real-World Data (RWD) team to identify patients with a disease prone to missed or delayed diagnosis.

The situation

A North American biotech company approached Clarivate™ about a data-driven project to identify patients with a specific rare, inherited condition prone to missed or delayed diagnosis. The company had a non-machine learning algorithm that used EHR data to find potential patients within its disease of interest, but struggled with its lower accuracy, and this approach missed many potential patients.

The need

Effectively employing data-driven methods, including machine learning, to identify suspect rare disease patients with improved accuracy.

Industry
Biotech

Therapy area
Rare disease

Clarivate™

How to avoid the pitfalls of a "garbage in, garbage out" scenario, biases, and inaccuracies?

If a machine learning algorithm is not fed the right data, it is not learning the correct signs to look for. The Clarivate team knew that laying the groundwork was as important to the project outcomes as the machine learning algorithm itself.

“We studied and characterized the cohort extensively with traditional methods to identify an appropriate control group that could minimize misclassification,” explained Hemanth Nair, Senior Director of RWD Product Management at Clarivate. There are diseases, he continued, that had similar symptoms to the rare disease the project was trying to identify patients for. “If you're not careful, and you don't explicitly put those patients as part of the comparison group and understand their profiles, you easily have a machine learning model that’s misclassifying a patient as a genetically based rare disease patient,” said Nair.

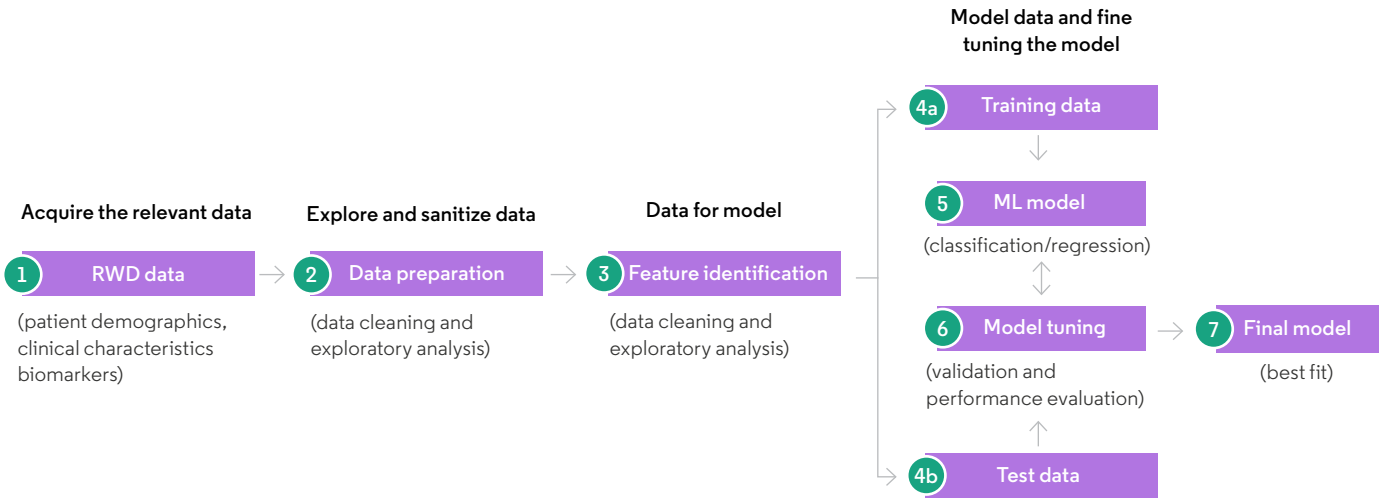
The team then embarked on feature selection, the process of selecting a subset of relevant variables, attributes, or predictors from the original dataset to improve model performance. The goal of feature selection is to reduce the dimensionality of the dataset by retaining only the most informative and discriminative features while discarding irrelevant or redundant ones. Once feature selection was complete, the model was ready to be trained, tested, tuned, and eventually validated. Because Clarivate has access to timely data, validation was able to be done on a set of patients that had the rare disease in question in 2023. The next step for the project, which initially took four months, is to forecast models to find patients before diagnosis, helping the patient, provider and payer to realize improved health outcomes.

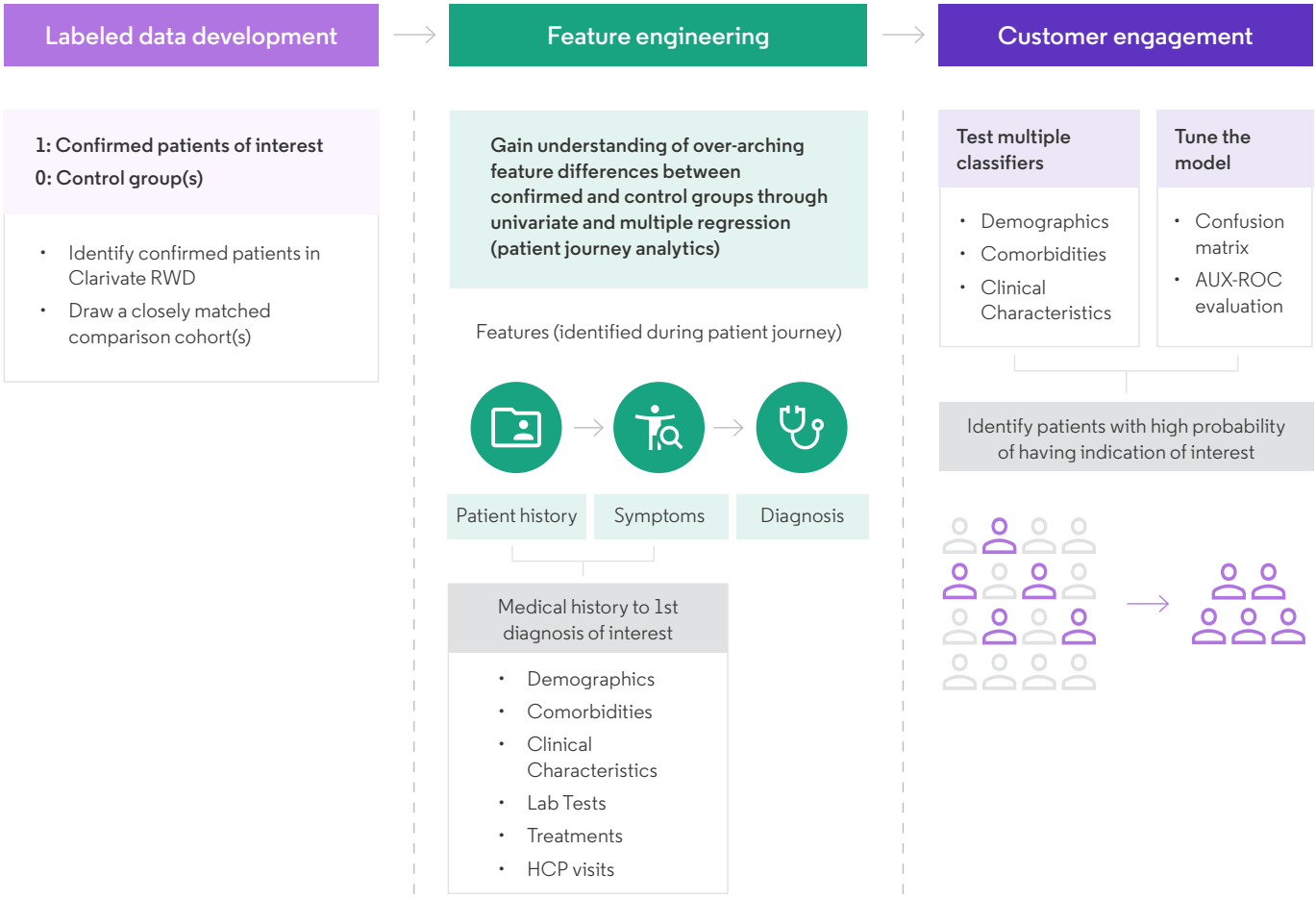
"There are diseases that had similar symptoms to the rare disease the project was trying to identify patients for. If you're not careful, and you don't explicitly put those patients as part of the comparison group and understand their profiles, you easily have a machine learning model that’s misclassifying a patient as a genetically based rare disease patient."

Hemanth Nair,
Senior Director of RWD Product Management at Clarivate

Clarivate ML team will leverage RWD and other data assets to train, tune and optimize models to predict patients of interest.

Figure 1: Machine learning approach





BXD1325786912 / 03

About Clarivate

Clarivate™ is a leading global provider of transformative intelligence. We offer enriched data, insights & analytics, workflow solutions and expert services in the areas of Academia & Government, Intellectual Property and Life Sciences & Healthcare. For more information, please visit **clarivate.com**.

Explore the way Clarivate Real-World Data fuels research breakthroughs:

clarivate.com/products/real-world-data/peer-reviewed-publication-highlights/

Contact our experts today:

clarivate.com