



Standardizing Clinical Data Labeling and Increasing Speed to Insights

Diaceutics data delivers the insights that medical experts need, with data sources including lab data, medical claims data, prescription data, and lab demographics. There are multiple data points including patient demographics, physician information, test results and reports, sample requirements, assay sensitivity, and many more. Stakeholders involved are patients, physicians, laboratories, and payers.

Company

Diaceutics is a data analytics and end to end services provider, and the world's largest repository of diagnostic testing data integrated in one secure platform. Diaceutics supplies consultancy services to some of the largest pharma companies.

Solution highlights

Standardized process, which is **scalable and shareable**.

Life Sciences Clinical Data
Data Labeling

Streamlining Data Analysis and Enabling Domain Expert Input

With so much data available and so many details within the data itself, an approach was needed to streamline the analysis of this data and empower analysts to add their medical knowledge to further enrich it. By using KNIME and taking advantage of all the tools, it's possible to cleanse and label the data and then fit it in a standard workflow for project-specific analysts to easily use. This saves them time and improves project quality. This specific example shows how Diaceutics implemented logic as well as business rules to label data, and how a standard workflow for project use was created.

Labeling of Clinical Data Allows Data-Driven Insights

There are several reasons why patient data needs to be labeled. Primarily, healthcare data is transactional. In this raw form, it offers little insight, from which no data-driven insight can be made. Labeling data appropriately allows insights to be uncovered and provides a cleaner and easier dataset to work with. It also allows the creation of groupings and filters. Not only for project-specific internal analysts to work with, but also for the Diaceutics DXRX platform, which clients can directly interact with themselves. The data that needs to be labeled varies and includes time point, disease, disease stage, patient history, biomarker tested, and test method.

For some of these, the task is to standardize or group the existing data. For example, with time points it's possible to group a specific data field by year, quarter, and/or month using a simple SQL statement. However, many parts of the data require a new label, created using a combination of logic and business rules - for example, disease stage.

In terms of labeling the data, straight-forward data can be hard-coded in SQL. However, for most data, control files and flexible SQL coding is used. In KNIME, linked components are used, which are files that contain all the logic for diseases: stage, biomarkers, methodologies, and business rules. A Build SQL Component builds out the SQL for all combinations specified in the control files, or the options that are chosen by the business analyst or on DXRX.

A project typically consists of taking patient level data and analyzing and aggregating it as appropriate for the client. The initial process involved approaching these on a project-by-project basis. With this method, workflows very quickly got complex and difficult to quality control. As a result, this was time consuming and difficult for business analysts to inherit and adapt as needed.

KNIME Success Story

Diaceutics
Better Testing, Better Treatment



"Since starting at Diaceutics, KNIME has been an integral part of my everyday work."

Isabel Stacey, Senior Data Analyst, Diaceutics

A Standardized, Six-Step Process

With a standardized approach, there is one agreed-upon method for all projects, and one way to do common client requests. This is easier to use, more consistent, and saves time. It also makes it easier for analysts to work independently and adapt when needed. In the workflow, there are only a few nodes that an analyst needs to interact with to complete their client requests. It keeps the full patient cohort aligned across projects with minimal quality control needed.

1. Connect to the database where patient-level data is located.
2. Pull out column names to create filter options in the menu.
3. Introduce control files using linked components.
4. Combine control files and columns from the original data table to create an interactive menu. Here, the user picks the options they need, the SQL code is automatically built out based on the combination the analyst has selected, and any business rules are built as variables for potential further analysis in the database query.
5. Variables from step four are implemented into query, and the analyst can choose to aggregate the data however they need.
6. Data is read out and ready for project team and client.

The Results: Better Data, Better Testing, and Better Treatment

This project has shown that it's possible to label healthcare data with many different variables including disease, disease stage, tested Biomarkers, method, and results. Labeled patient data and a standardized process ensures all analysts are working from the same base. Anyone working with the data has the same starting point with same patient cohort and methods. This means data can be analyzed and aggregated at a high level quickly and efficiently. In depth analysis can be performed more easily (when needed) as the patient cohort is readily available. This ultimately leads to better data, better testing, and better treatment.

Why KNIME?

KNIME workflows are easy to build and allow a straightforward way to standardize business processes. All nodes and sections can be annotated, which not only provides a self-documenting workflow, but enables a new user to understand what is happening at each stage. One of the biggest benefits of KNIME is the linked component functionality. This ensures that changes can be made to the master workflow, and all versions of that downloaded workflow will get a notification warning the user that a change has been made. This also enables version control of the workflows, by using snapshots on the KNIME Server.

From a business perspective, this solution has highlighted how easy it is to scale with standardized workflows, without the risk of having different analysts interpreting different results out of the workflows. With the evolution of the data lake, KNIME, and ETL processes, project throughput has increased significantly – specifically moving from an Excel-based approach to this more standardized, streamlined approach.