Uncovering Critical Breast Cancer Insights Hidden in EHR Notes

Client: Oncology-Focused Consulting and Real-World Evidence Firm

Project:

- AI-powered EHR dataset researching precise patient-level insights on cancer stage and HER2 status from clinical notes.

Challenge:

- Cancer staging and HER2 status are **not stored in structured fields, and clinicians describe them in highly variable**, vague ways across notes.

egnite

Approach

- egnite's data science team developed a **multi-layered extraction system combining traditional NLP techniques with large language models (LLMs)** to decode the nuanced nomenclature and implicit references common in oncologist notes.

How We Did It

Unstructured Clinical Note: *"...prev. external eval listed T2N0, now T3..."*

egnite Algorithm: Stage 3

Unstructured Clinical Note: "...familial history of stage iv bc, mother & aunt, patient stage ii..."

egnite Algorithm: Stage 2

Unstructured Clinical Note: *"...microinvasive ductal, < 1mm, grade II intraductal 1.4 cm, N0, ER/PR+, pos Her2..."*

egnite Algorithm: Positive

Impact

- Within 30 days, egnite **delivered a fully de-identified, research-grade dataset** of breast cancer patients, each with **a confirmed cancer stage (0-4) and HER2 status (positive/negative)** all extracted from free-text notes.
- By transforming previously inaccessible unstructured data into **structured, actionable insights**, the client intended to accelerate critical development decisions with speed and confidence.

LEARN MORE ABOUT EGNITE'S RWD