CASE STUDY

Real-Time Data Quality Assessment with Elucidata's Polly

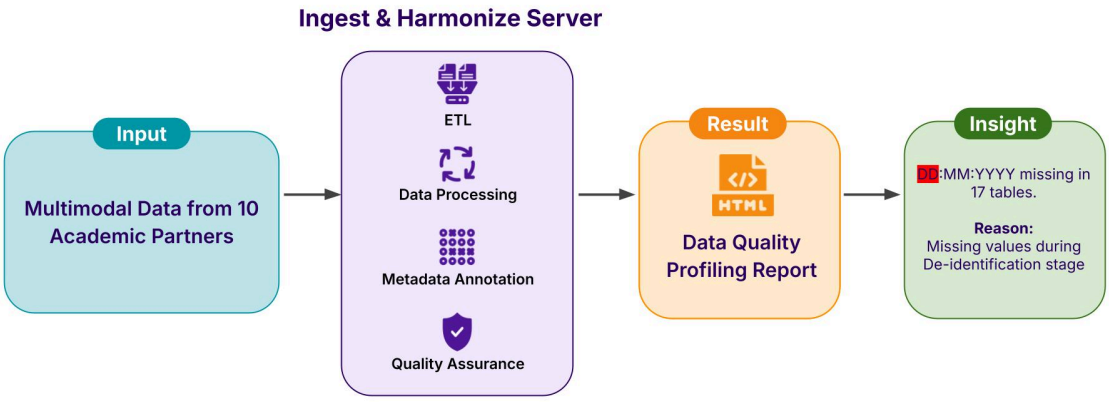
A large diagnostics company wanted to accelerate new product development in Hospital-acquired Sepsis by integrating **EHR, imaging & sequencing data** from multiple partners. The incoming datasets **(~30M patient records)** were highly heterogeneous, with inconsistencies, missing values, and structural discrepancies. They also had to be standardized to OMOP, the company's chosen data model. Wrangling this massive volume of data without automation was nearly impossible. They used **Elucidata's Polly** to flag low-quality datasets and transform them into AI-ready data, saving **years of manual assessment and curation efforts.**

Problem Statement

- A team faced the challenge of harmonizing millions of patient records from **10+ academic partners and vendors**, each with different formats and metadata. This complex multi-modal data included **30M EHR data, 50M imaging records, and 25M omics datasets.**
- Data heterogeneity was a major obstacle, with **fragmented records and non-standardized terminologies** complicating efforts to establish provenance, completeness, and reliability.
- **Without a scalable, systematic approach to data quality assessment**, downstream analyses were compromised, delaying new product development & R&D in Sepsis.

Solution: Data Quality Assessment by Polly

Polly, Elucidata's **AI-ready clinical data platform**, is designed to ensure high-quality R&D data across diverse sources and more than **25+ modalities**. Its multi-modal data model integrates data from EHR, imaging, omics profiles, and clinical trials. Additionally, Polly's ingest module includes a built-in data validation mechanism to ensure **conformance, completeness, consistency,** and **plausibility** in incoming datasets.



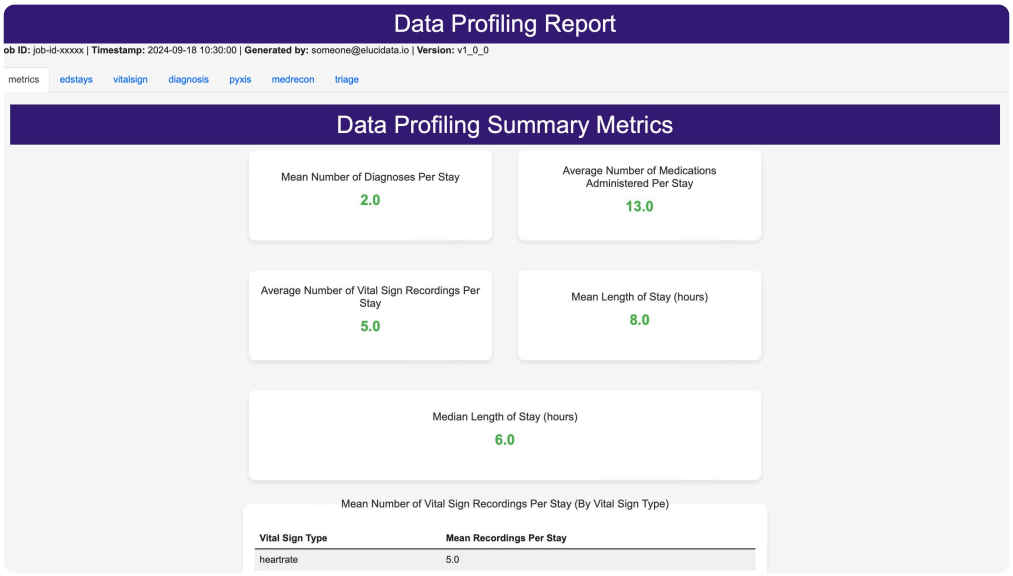
Data Quality Assessment In-Real Time With Polly

Data Ingestion and Profiling for EHR, Imaging & Sequencing data

Polly's data quality assessment starts with **data ingestion and profiling** to ensure high-quality multi-modal data integration by identifying inconsistencies, missing values, and format deviations. This initial check assesses field distributions, datatype variations, and anomalies, evaluating data for:

- **Completeness Metrics:** Analyzes gaps in time-related fields (e.g., missing timestamps in `vitalsign.csv, edstays.csv),` and checks for missing diagnoses `(diagnosis.csv)` or medication `(medrecon.csv, pyxis.csv).`
- **Consistency Metrics:** Ensures consistency by verifying matching `stay_id` values across tables and validating the uniqueness of `stay_id` in `edstays.csv.` It also checks ICD codes for correct formats using regex validation.

Following this, **pre-harmonization QC reports** validate format adherence and relational integrity. Automated validation mechanisms detect and flag errors, inconsistencies, and missing values in raw data by identifying duplicate records, missing attributes, incorrect timestamps. Then, this unstructured data is standardized and harmonized to ensure compliance with industry-standard models such as OMOP CDM.



Quality Assurance and Data Harmonization

This published **pre-harmonization QC reports** identified gaps, inconsistencies and discrepancies in the patient data, which were addressed through the data harmonization process to improve data quality, and usability.

One of the major challenges identified in the reports was **date-time inconsistencies.** The required date time format was `yyyy-mm-dd hh:mm:ss` but the source data contained several issues, such as missing day values, incorrect time formatting, and a wide range of years across datasets. For example, in the table below **Problem_List_2022** dataset had years ranging from `1888 to 2023,` while the **Procedures_2022** dataset included entries from `1901 to 2048.`

Table Name	Field Name	Observed format	Expected format	Values observed
Encounter_2022	EncounterStartDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	yyyy - 2021, 2022, 2023 mm - 1,2,3,4,5,6,7,8,9,10,11,12 **dd - absent** **hh - < 12 (not in 24 hr format?)**
Encounter_2022	EncounterEndDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	yyyy - 2021, 2022, 2023 mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)
Problem_List_2022	ProblemStartDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	**yyyy - range (1888 to 2023)** mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)
Vital_Signs_2022	ObservationDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	yyyy - range (2002 to 2023) mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)
Medications_2022	MedicationDocumentationDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	yyyy - 2023, 2022 mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)
Procedures_2022	ProcedureEffectiveDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	**yyyy - range (1901 to 2048)** mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)
LaboratoryResults_2022	ResultDate	yyyy-mm hh:mm:ss	yyyy-mm-dd hh:mm:ss	yyyy - 2019, 2021, 2022, 2023 mm - 1,2,3,4,5,6,7,8,9,10,11,12 dd - absent hh - < 12 (not in 24 hr format?)

Another significant challenge flagged in the reports was the **presence of missing data.** The analysis revealed several fields with over 75% missing values, which were categorized as "Extremely High Missingness." For example, in the table below metadata fields like **PatientRaceText** exhibited missingness rates exceeding 85%, while **EncounterDiagnosisCodeSystem** also showed substantial gaps, making it a crucial area for data harmonization interventions.

*This report shows fields (Field Name), across tables, with >75% values missing ("**Extremely High Missingness**" category) Other fields, across tables, dont have very high missingness.*

Table Name	Field Name	Count of missing values	% of missing values
Patient_2022	PatientRaceText	8648348	85%
Patient_2022	PatientEthnicityText	10028381	99%
Problem_list_2022	ProblemText	20627794	99.78%
Diagnosis_2022	EncounterDiagnosisCodeSyster	32554366	93.82%
Encounter_2022	HR_24_FLAG	9962608	97.03%
Insurance_2022	InsuranceCompany	3111401	77.46%
LaboratoryResults_2022	ResultDate	327597811	96.00%

All missing values are "N". To be used for **visit ID selection** in absence of datetime information

We harmonized and standardized missing day components for **293 fields to a 24-hour format** and aligned timestamps across key metadata categories. AI-powered anomaly detection was employed to flag critical gaps in data integrity, and schema modifications were recommended to mitigate systemic data loss.

High-Quality Data With AI-driven Insights

To ensure data integrity, a **post-harmonization QC report** is generated with **99.99% comprehensive data coverage**, confirming that no information is altered, lost, or misrepresented during transformation. This robust quality check is seamlessly integrated with AI-powered cohort builders, enabling users to generate no-code data insights that are easily shareable and reusable with collaborators. This allows users to generate **real-time insights through AI-powered analytics and with quality validation.**

Impact

Opportunity to accelerate new product development by **25%**

6X faster Data Product Creation & Analysis

Reduced data management and data operations costs by **5M/Yr**

4X cheaper Multi-modal Data Products Generation with Robust Quality Control at each step

About Elucidata

Elucidata's Polly optimizes data quality for pre-clinical and clinical drug discovery by harmonizing multi-omics and assay data into ML-ready formats. Polly's powerful harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable costs.

Polly's advanced technology accommodates 25+ R&D data types, catering to teams in pre-clinical drug discovery and diagnostics R&D. Polly is trusted by over 25 research organizations, including 4 of the top 10 pharma companies, to accelerate their discovery programs.

Book a Demo