CASE STUDY

Elucidata Scales Clinical Data Extraction from Hospital Records 8x Faster than Manual Methods

A leading Indian hospital group specializing in advanced medical care sought to optimize clinical operations, enable predictive analytics, and develop personalized treatment programs by leveraging patient records. This required digitizing and harmonizing a vast corpus of 350,000 clinical records—including handwritten notes, PDFs, and scanned images—along with integrating molecular test data.

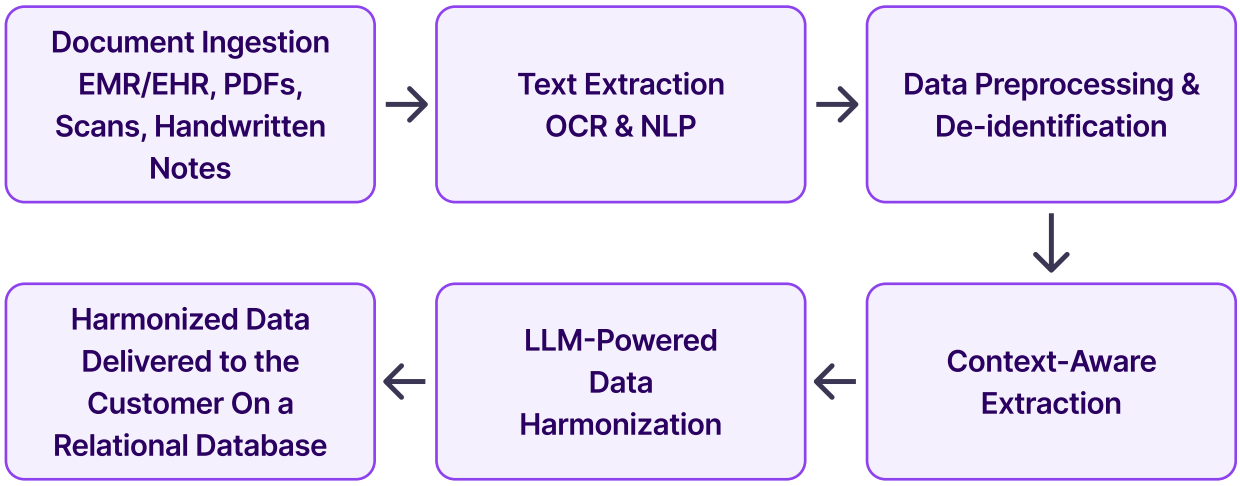
Given the scale, **manual transcription was impractical.** The clinical operations team needed a scalable, **automated approach to digitize, clean, and map the data** to critical metadata using standardized ontologies. To address this, they partnered with Elucidata to build & deploy a production-grade AI pipeline with Elucidata's data-centric AI platform, **Polly**. The pipeline leverages LLMs and data processing components to deliver accurate, secure, and interoperable clinical datasets **8x faster** than manual methods.

Challenge

- **High Data Volume & Complexity:** The organization was handling thousands of clinical documents from various sources, each with a different layout and format.
- **Manual & Inefficient Workflows:** Traditional manual transcription of clinical documents was time-consuming, error-prone, and not easily scalable.
- **Limited Standardization & Interoperability:** Harmonizing data with established clinical vocabularies (e.g., ICD-11, ICD-10, SNOMED CT, LOINC, RxNorm) was critical to ensure interoperability across clinical research, regulatory, and decision-making.
- **Regulatory Compliance:** Handling sensitive patient data required robust compliance measures and secure data processing frameworks.

The Solution

Elucidata utilized it's Data-Centric platform, Polly to develop AI-driven, scalable pipelines that extract, harmonize, and structure medical information from unstructured text at scale. The pipeline linked data from EHRs, PDFs, free-text notes, scanned images, and molecular tests, creating a comprehensive view of individual patients and cohorts. The final dataset was integrated, structured, enriched with metadata, and ready for AI applications.

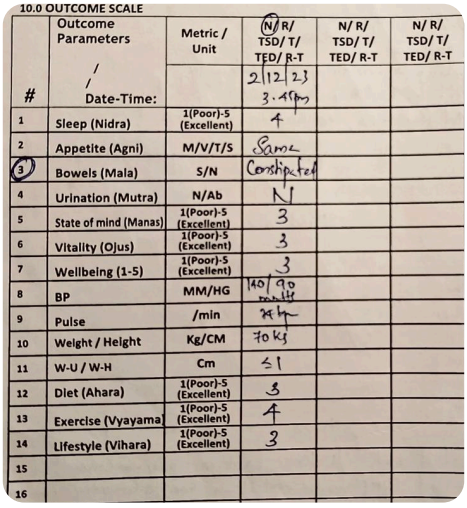


1. **Document Ingestion:** Clinical documents in various formats—EMR/EHR text, PDFs, scans, and handwritten notes—were ingested and automatically classified by type (e.g., lab reports, discharge summaries, radiology notes) using Polly's Ingestion Workflows. The data ingestion and profiling system routed each document type through a specialized pipeline to ensure optimal extraction accuracy.
2. **Text Extraction (OCR & NLP):** For non-readable PDFs and handwritten notes, AI-based OCR models converted images into structured text. By leveraging advanced AI-based vision algorithms, Polly accurately captured typed and handwritten text, tables, and forms with minimal configuration.
3. **Data Preprocessing & De-identification:** The extracted text was systematically de-identified using automated tools to remove personally identifiable information (PII), ensuring compliance with HIPAA and GDPR standards. This step enabled privacy-preserved downstream analytics while safeguarding patient confidentiality.
4. **Entity Extraction & Concept Recognition:** Polly's LLM Models, fine-tuned for clinical use cases, identified critical entities such as diagnoses, medications, lab results, demographics, and disease states. Context-aware extraction captured dosage information, temporal references, and clinical negations, enhancing prediction accuracy.
5. **Data Harmonization:** Extracted entities were also harmonized using these models. Metadata fields across 16 categories (e.g., demographics, diagnoses, family history, allergies) were mapped to industry-standard terminologies like ICD-11, SNOMED CT, LOINC, and RxNorm.

   With support for over 1,200 attributes, **Polly's Unified Data Model** assured seamless data integration while remaining adaptable to new samples, metadata, and modalities. This removes duplicates, consolidates repeated data points from multiple documents, and organizes each concept within a relational schema designed for flexible analytics and compliance.
6. **Storage & Access:** The processed data was stored in a secure, high-performance relational database on Polly, optimized for rapid querying and real-time analytics. Controlled access was provided through user-friendly dashboards and secure APIs, enabling seamless integration with clinical workflows, supporting research, and personalized treatment development.

Outcome

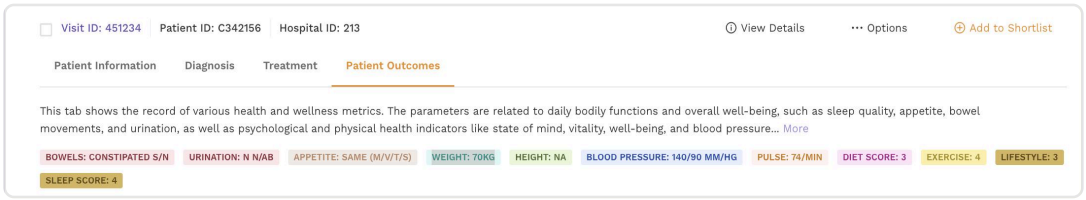
LLM-powered metadata extraction from patient records



10.0 OUTCOME SCALE

#	Outcome Parameters / / Date-Time:	Metric / Unit	(N)/ R/ TSD/ T/ TED/ R-T	N/ R/ TSD/ T/ TED/ R-T	N/ R/ TSD/ T/ TED/ R-T
			2/12/23 3.45pm		
1	Sleep (Nidra)	1(Poor)-5 (Excellent)	4		
2	Appetite (Agni)	M/V/T/S	Same		
3	Bowels (Mala)	S/N	Constipated		
4	Urination (Mutra)	N/Ab	N		
5	State of mind (Manas)	1(Poor)-5 (Excellent)	3		
6	Vitality (Ojus)	1(Poor)-5 (Excellent)	3		
7	Wellbeing (1-5)	1(Poor)-5 (Excellent)	3		
8	BP	MM/HG	140/90 mmHg		
9	Pulse	/min	74bpm		
10	Weight / Height	Kg/CM	70 kg		
11	W-U / W-H	Cm	≤1		
12	Diet (Ahara)	1(Poor)-5 (Excellent)	3		
13	Exercise (Vyayama)	1(Poor)-5 (Excellent)	4		
14	Lifestyle (Vihara)	1(Poor)-5 (Excellent)	3		
15					
16					

Handwritten Patient Record **(Before)**



Digitized File **(After)**

Impact

8X Faster — **Accelerated text extraction compared to manual transcription**

5X Cheaper — **Delivered a cost-effective solution for text extraction**

About Elucidata

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

Elucidata's advanced technology accommodates 25+ R&D data types, supporting teams in pre-clinical drug discovery and diagnostics R&D. Elucidata is trusted by leading pharmaceutical and biotechnology companies to accelerate their discovery programs.

Book a Demo