John Snow LABS

phuse connect
connect·share·advance

Scaling Regulatory-Grade RWE: A Hybrid NLP, SLM and Deterministic Reasoning Framework for Automated Cancer Registry Abstraction

Understanding Cancer Registries

What is a Cancer Registry?

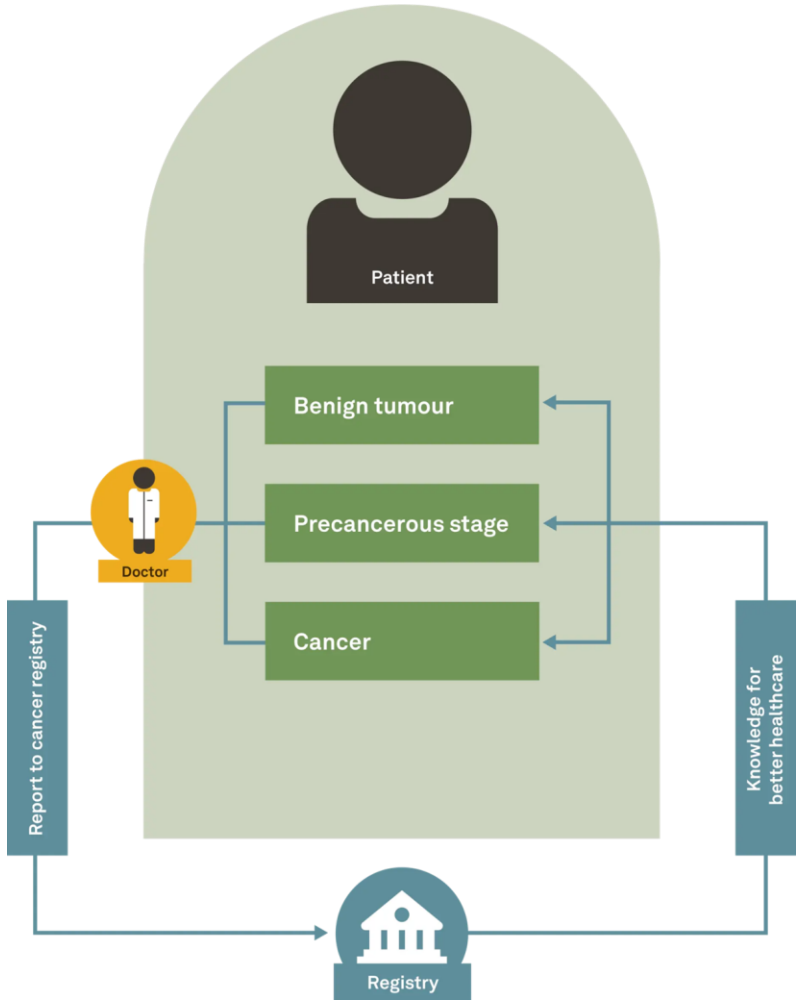
A systematic collection of data about cancer patients, including demographics, diagnosis details, treatments, and outcomes. These registries serve as **critical resources for monitoring cancer patterns** and improving patient care.

Why Are They Needed?

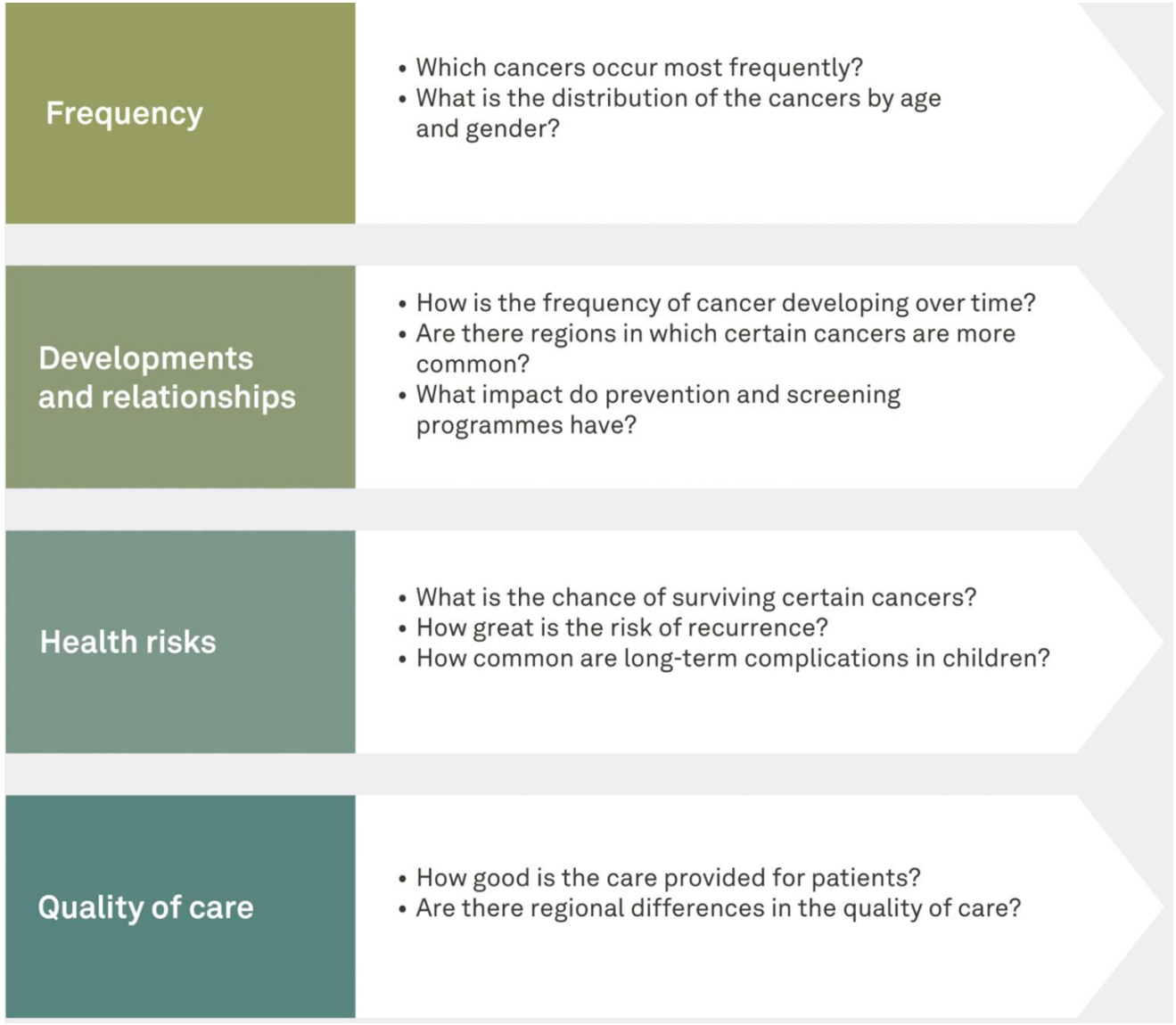
Cancer registries enable monitoring of incidence rates, survival trends, and treatment effectiveness. They inform **public health policies**, support clinical research, ensure quality of care, and facilitate international collaboration.

The Traditional Process

Certified Tumor Registrars (CTRs) manually abstract data from thousands of pages of medical records, pathology reports, and imaging studies. This labor-intensive process takes **~2 hours per case** and faces significant backlogs.



What Cancer Registries Can and Cannot Answer



The typical delay from diagnosis to national reporting is **12-24 months**.

This means that cancer registries **cannot** not play a part in treating active patients:

- Clinical decision support
- Clinical trial matching
- Clinical guideline adherence
- Active care coordination
- Precision medicine interventions
- Early intervention for at-risk patients
- Immediate monitoring of events

The Burden of Manual Cancer Data Abstraction

NPCR timeliness target:	**90% within 12 months**
Registries meeting standard:	**Only 14%**
Average reporting lag:	**23 months nationally**
Backlog for processing:	**Up to 7 months**

2 hrs Average manual abstraction time per case

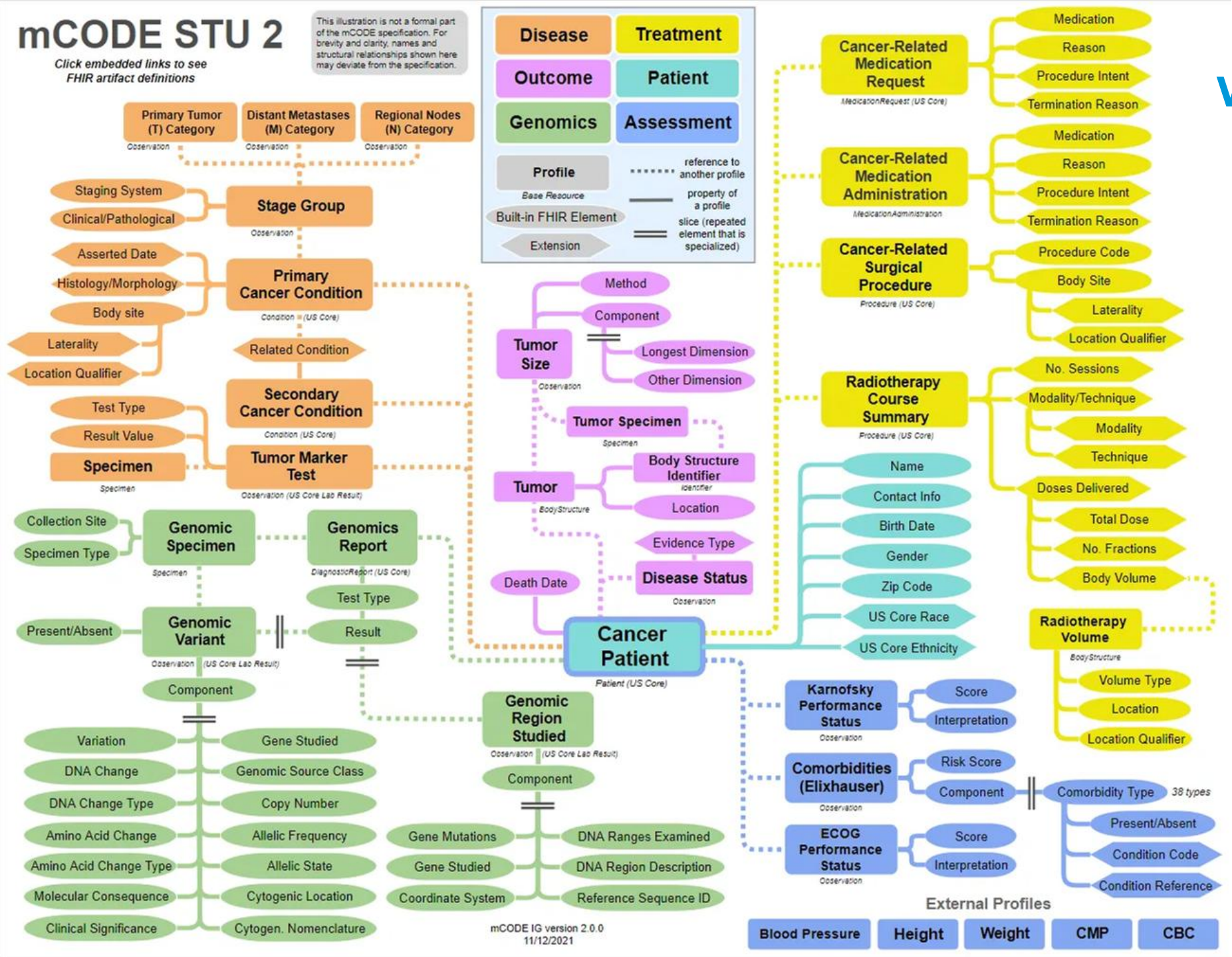
441 New cases annually effectively handled by a single full-time cancer registrar (Oncology Data Specialist)

$60-90K Annual salary range for a certified registrar

Document Complexity & Volume

Registrars must extract structured data from **thousands of pages** of unstructured clinical notes, reports, and images.

- Pathology Reports
- Radiology Images
- Clinical Notes
- Surgery Reports
- Genomic Tests
- Treatment Records



What's in a Registry?

- As of 2026, the NAACCR data dictionary contains **over 900 data fields**
- For a typical adult case, a registrar usually completes **200–250 required fields**
- Only about **40% of fields** can be automatically populated from an EHR
- Registrars must write **narrative summaries** for every major field to justify their coding.

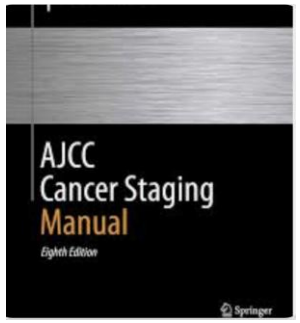
AJCC, SEER, SEPAR Guidelines:
2,500+ pages of text, flowcharts and tables

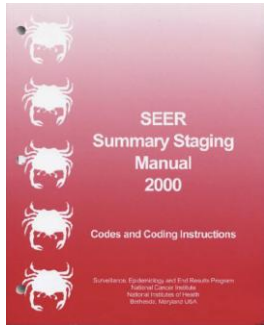
SEER & AJCC & SEPAR Guidelines: **2,500+ pages** including branching decision trees and cancer-specific taxonomies

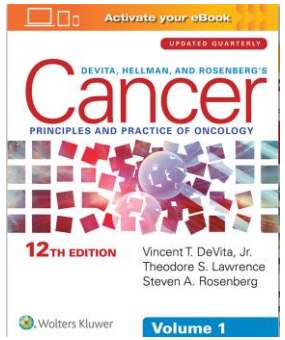
Complex Decision Logic: TNM staging requires integration of multiple data points across documents

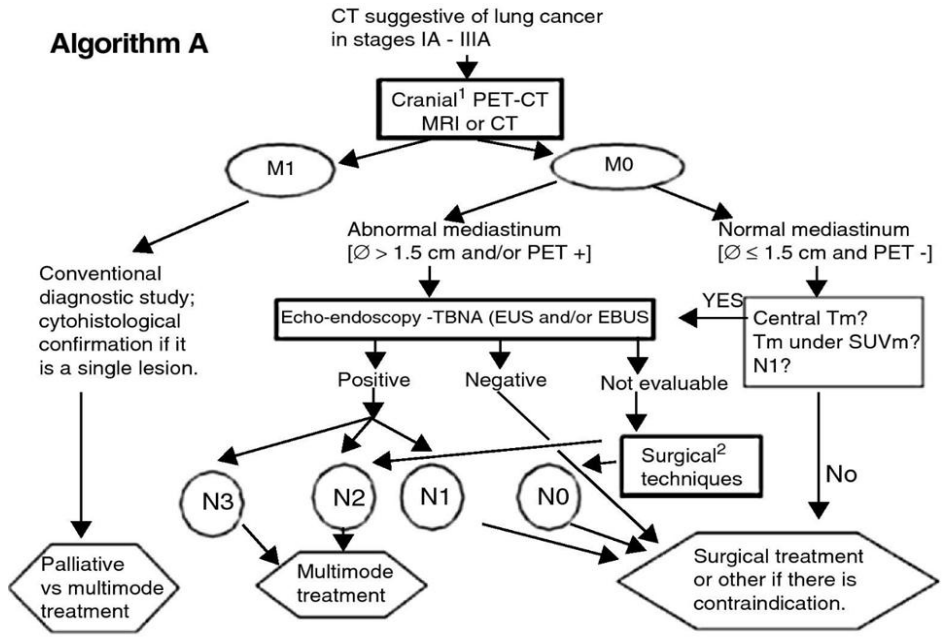
Cancer-Specific Rules: Each cancer type has unique staging, grading, and biomarker requirements

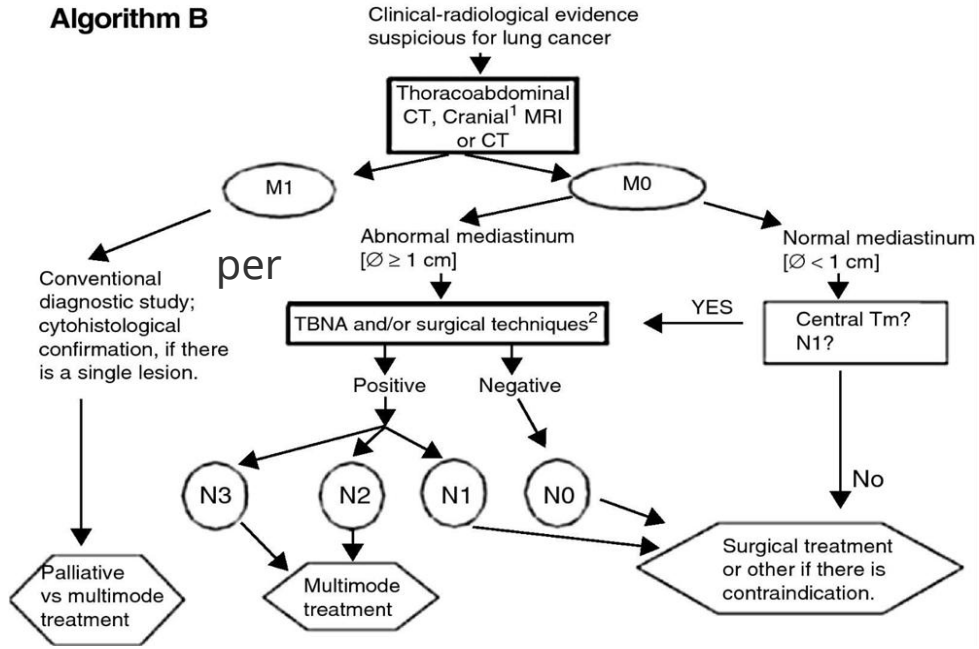
Evolving Standards: 2018 specification changes **doubled abstracting time per case**











John Snow Labs' approach to automating cancer registries combines **multimodal AI, agentic workflows**, and **human oversight.**

Data Curation & Extraction

Multimodal AI enables intelligent curation of cancer registry data by unifying diverse clinical sources – pathology reports, radiology findings, genomic tests, and physician notes – into structured, standardized formats. Using domain-tuned NLP and LLM models, the system automatically extracts key fields while applying deterministic rules to ensure consistency.

Patient-Level Reasoning

Through **agentic workflows and auto-consolidation,** the system automatically reconciles partial or modified records in real time, enabling staff to focus on “review by exception” instead of manual data consolidation.

Oncology-Specific Agents

Selecting guideline logic by cancer type; translating SEER/AJCC flowcharts into executable decision graphs for staging, histology, biomarkers, and treatments and compiling results in a digestible format thru certain ontologies such as mCode.

UI for Validation & Feedback

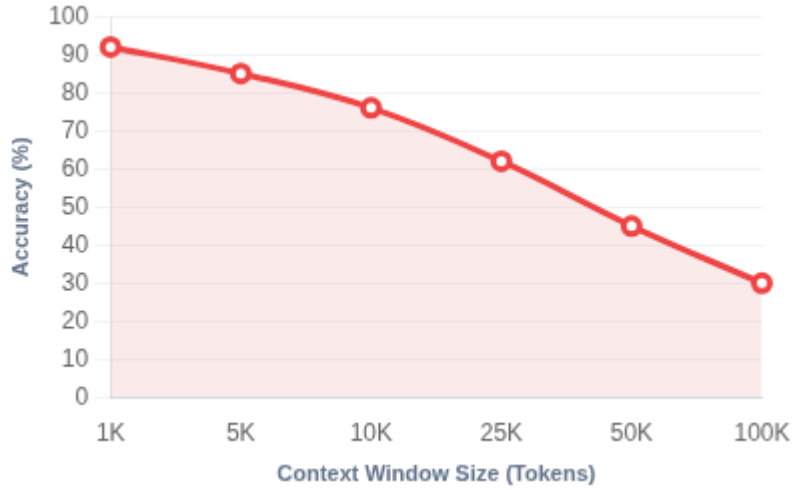
Workflows that registrars and auditors can use along with features for team collaboration, audit trails, and a learning loop that improves models over time.

What Doesn’t Work: “Dump it all into an LLM”

Naïve LLM-based strategies fail fast:

- 1M context window holds about 1,000 pages
- That’s less than one cancer patient’s data
- … and less than just the guidelines
- … discounting images and flowcharts
- The effective context window is much smaller

The "Context Rot" Phenomenon



Performance degradation in information retrieval as context window size increases (simulated based on Chroma/RAG benchmarks)

Regulatory Rigor

2,500+ Pages of Guidelines

- ✕ Volume of SEER/AJCC rules overwhelms human registrars & LLMs.
- ✕ Without specialized reasoning layers, consistent rule application is impossible.

Multimodal Data

5,000+ Pages per Patient

- ✕ Requires integration of pathology (text), radiology (reports), and labs (tables).
- ✕ Single-modality models miss cross-referencing cues essential for accurate staging.

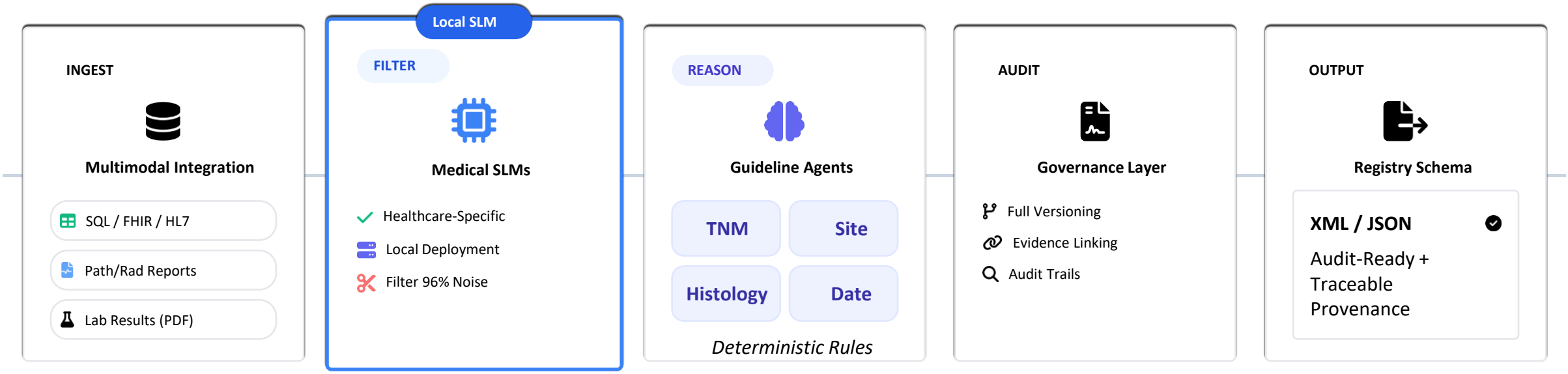
Deterministic Inference

- ✕ Patients have multi-year journeys with evolving disease states.
- ✕ Must generate the same answers every time: deterministic and reproducible answers

Auditability

- ✕ Black-box outputs are rejected by registries requiring provenance.
- ✕ Must link every variable to specific text spans for registry verification.

High-Level Solution Architecture: Delivering Accuracy, Scale, and Governance



Data Security

Local deployment eliminates egress risks; sensitive IP never leaves secure environment.

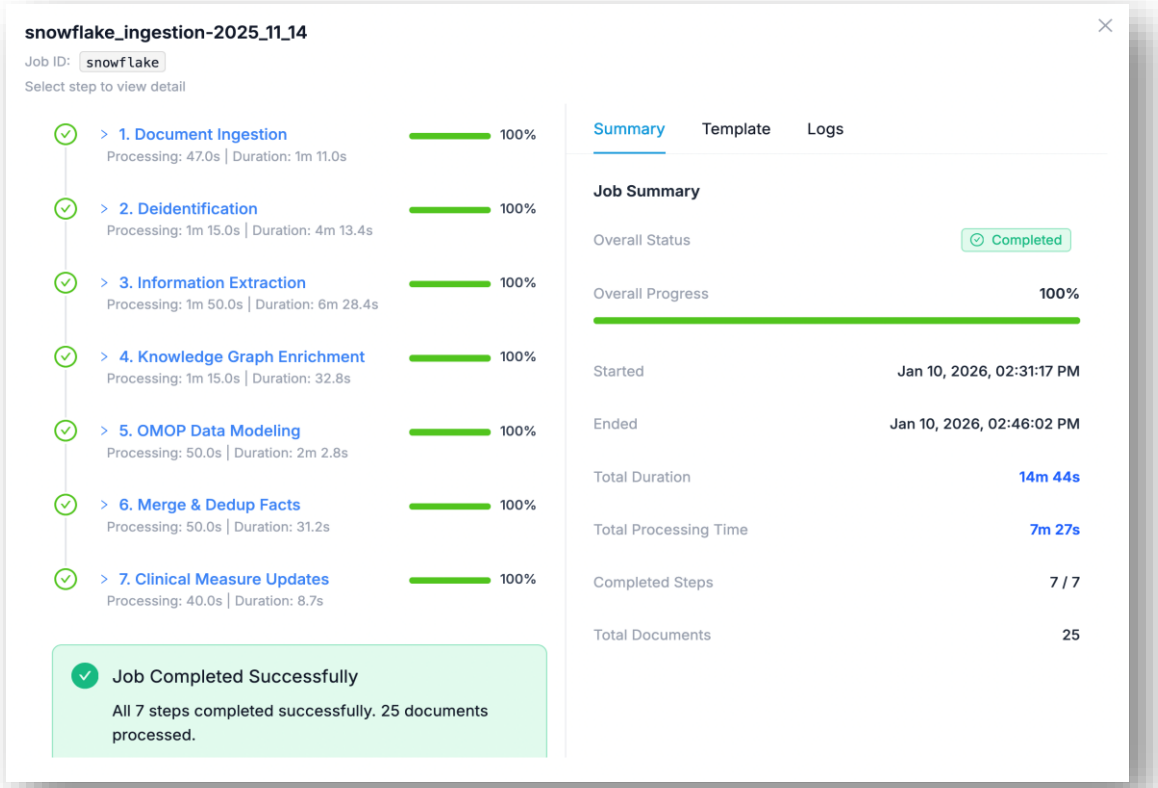
No API Egress

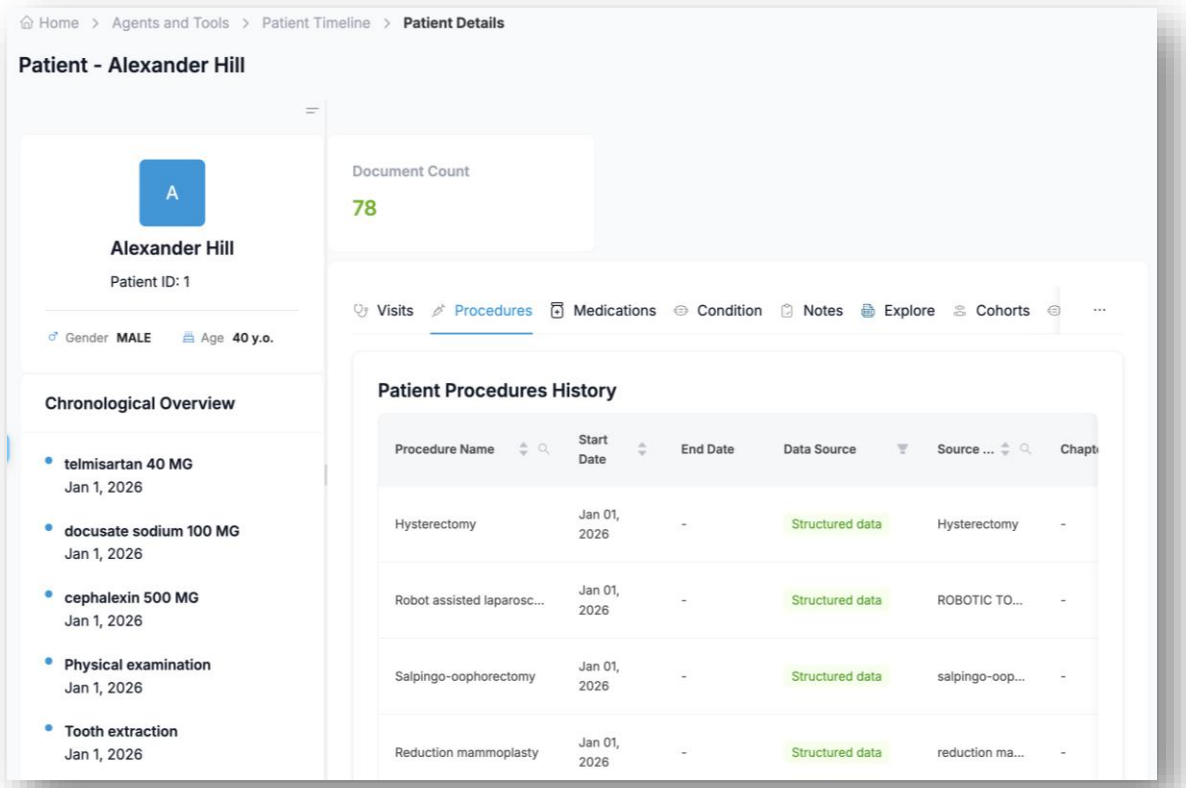
Bypasses high operational costs and privacy risks of frontier LLM APIs.

Governance-by-Design

Every extraction linked to source line; supports full data and model versioning.

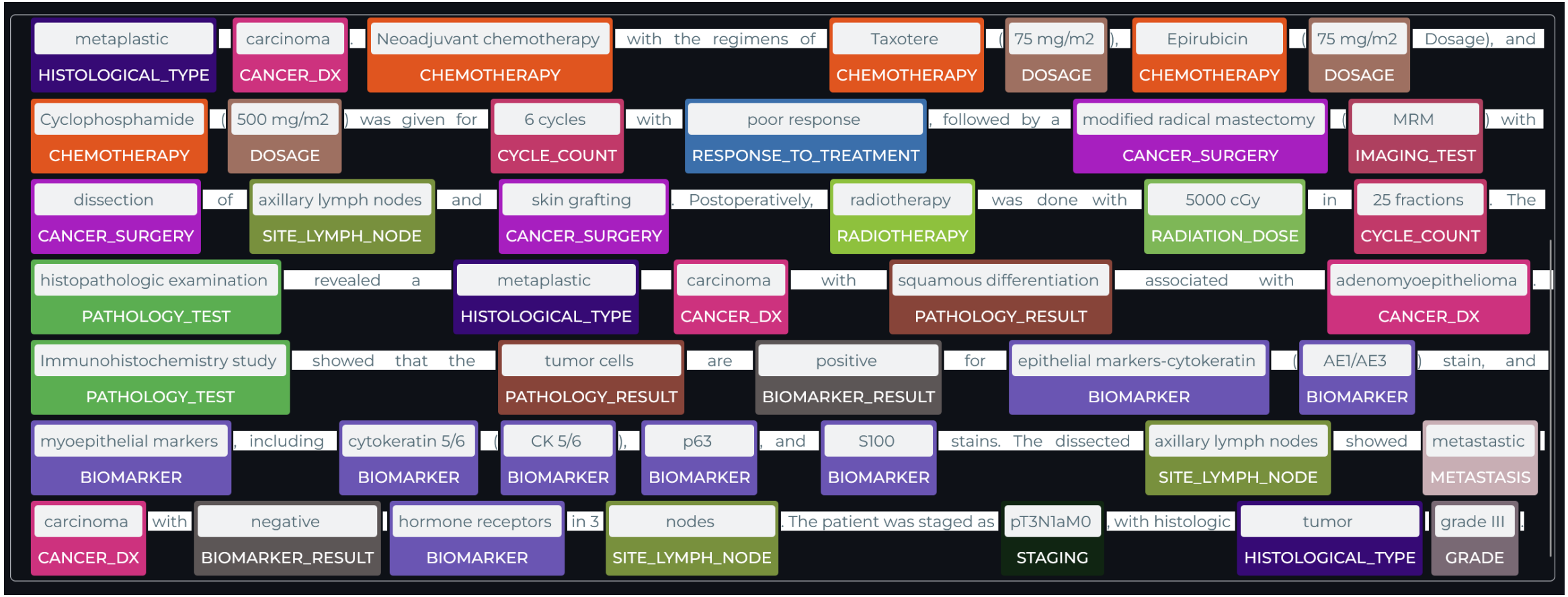
Multimodal Data Integration





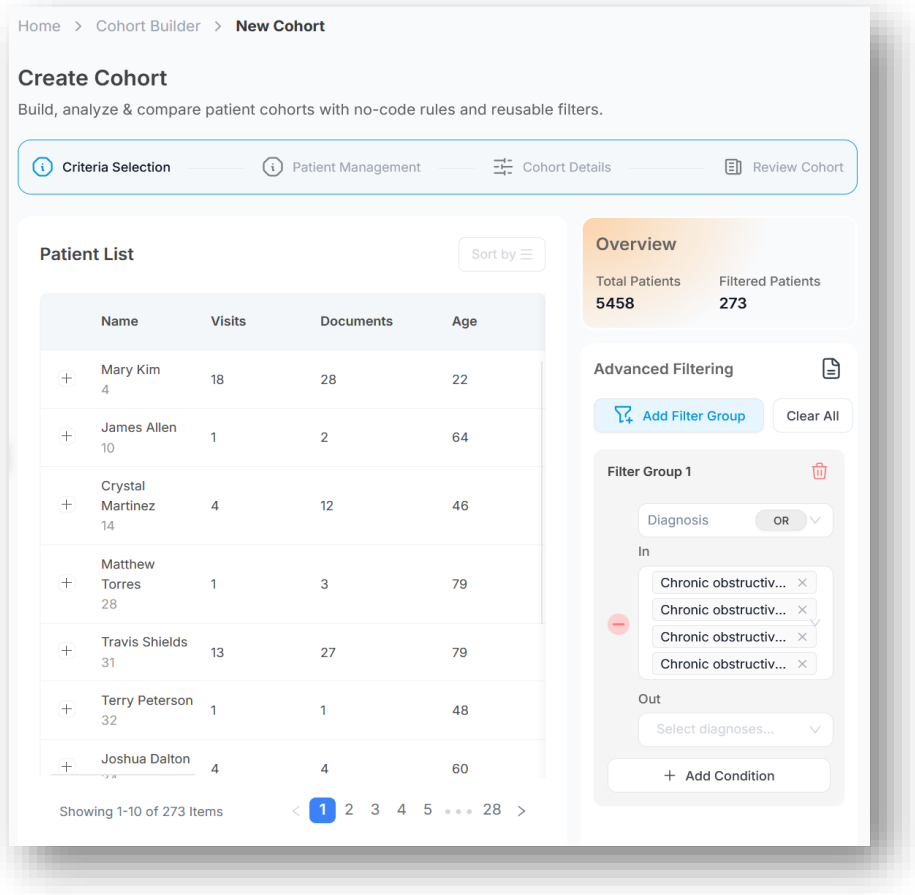
- Ingest and understand EHR, FHIR, Text, DICOM, and PDF (forms / scans)
- Extract information from visual & text documents and normalize terms to codes
- Transform the data to the OMOP data model, while merging & de-duplicating facts

Information Extraction with Clinical NLP



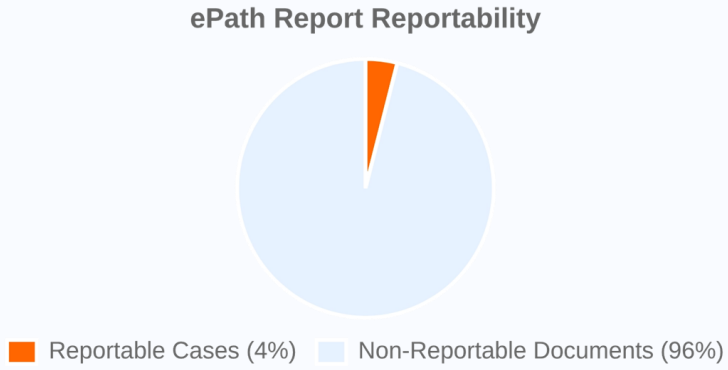
- John Snow Labs has built 3,000+ small, task-specific medical language models
- They are deterministic, making it easier to reproduce the exact version of a data pipeline
- They are far faster & cheaper for inference (5k pages of in & out tokens = $112.5 on Claude 4.6 Opus)

Case Finding



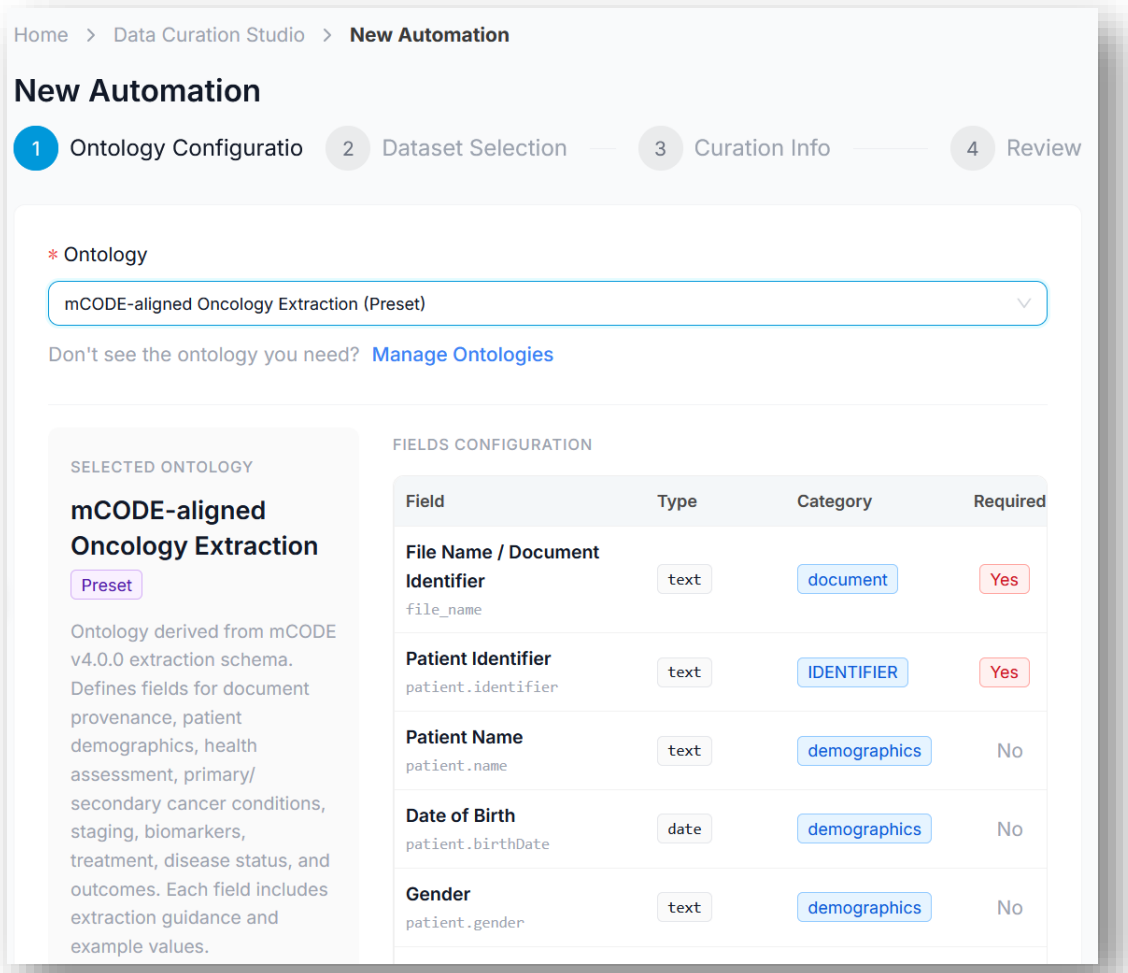
Signal-to-Noise Ratio

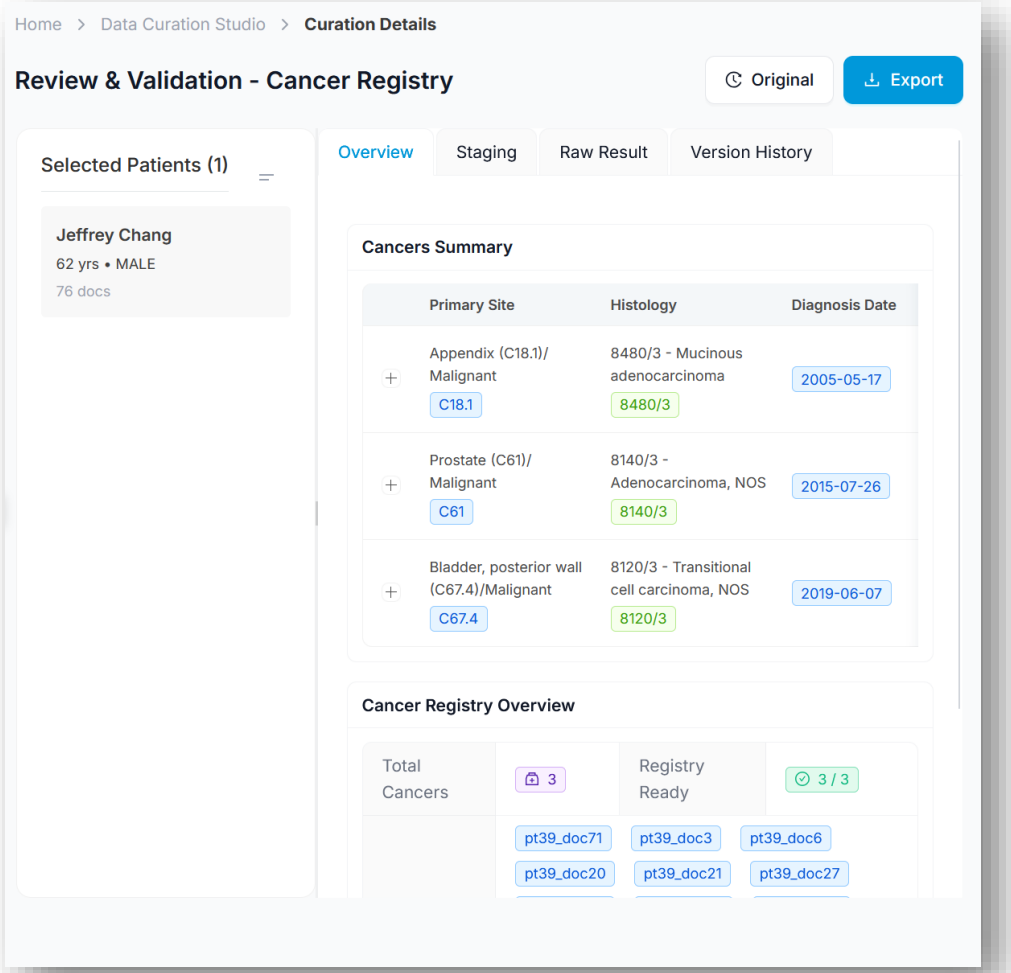
Only **4%** of electronic pathology (ePath) reports are actually reportable cancer cases.



- Case finding – finding the right patients to report on – takes 15%-25% of a registrar's time.
- The system identifies which patients are reportable, automatically and continuously.
- It then filters which documents are relevant for curation – saving LLM tokens and time.

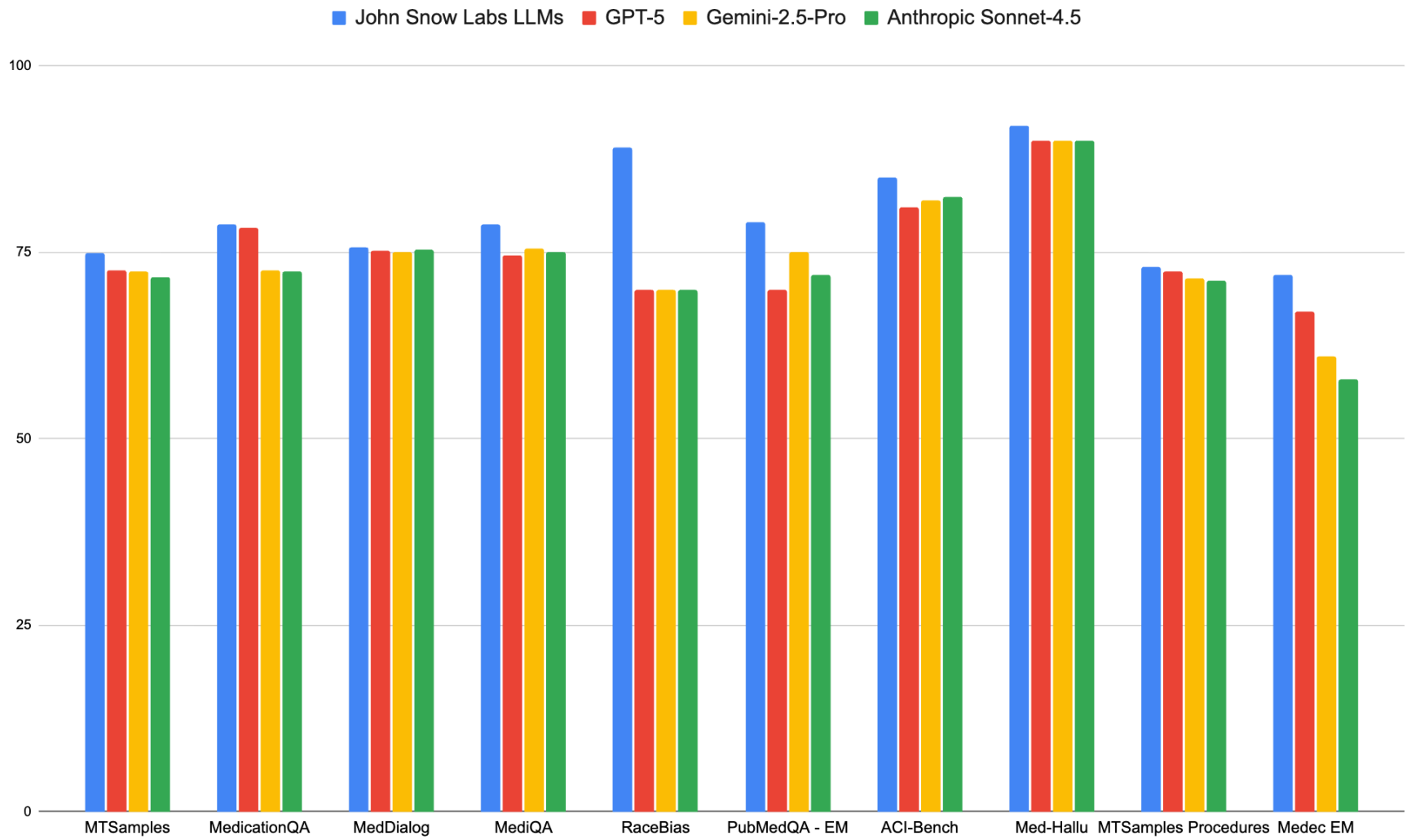
Automated Data Abstraction





- Automate the entire registry abstraction process, using the selected ontology & guidelines.
- Enable human-in-the-loop review, including full data provenance of changes made.

Medical LLMs Outperform General Frontier LLMs



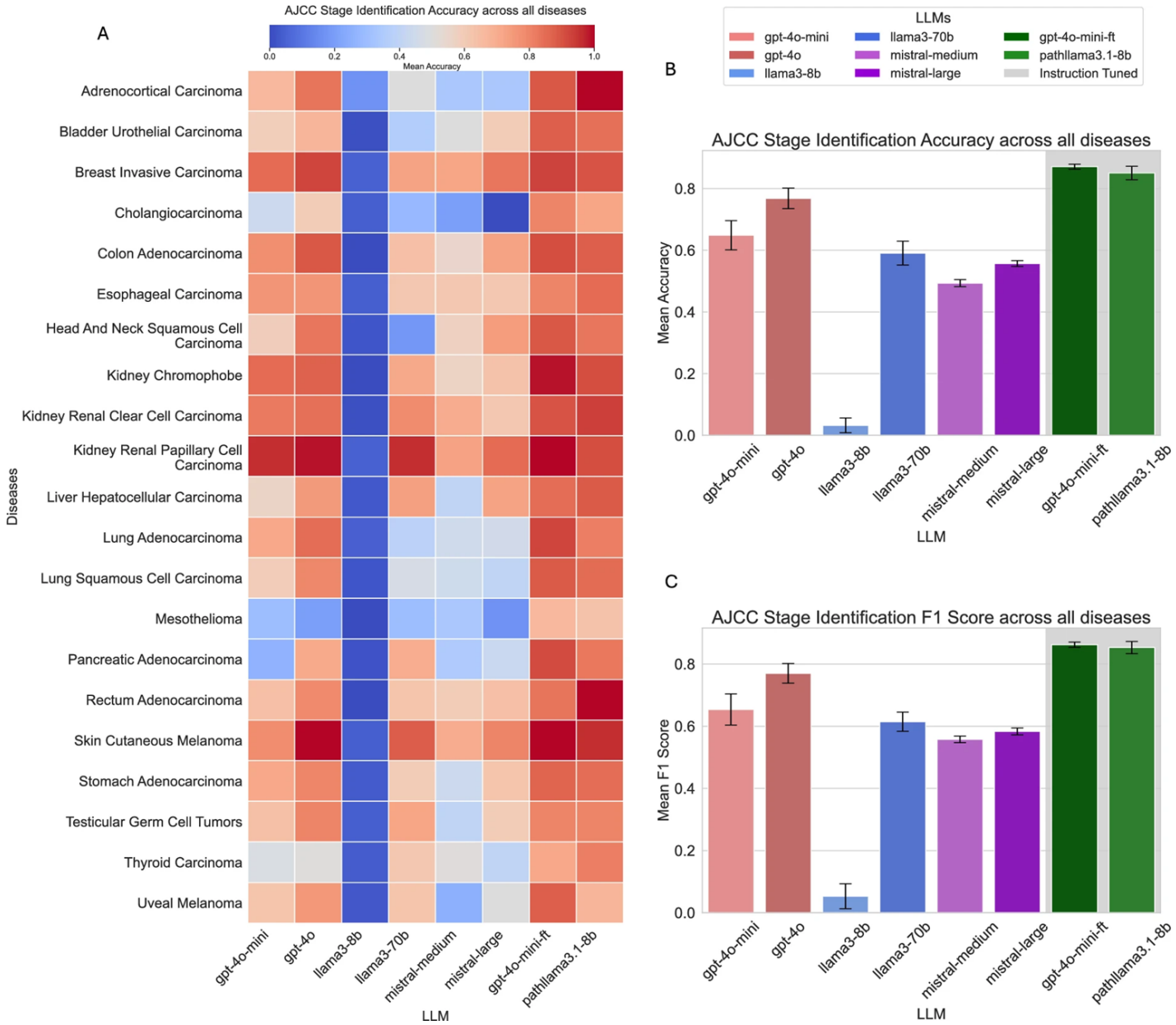
John Snow Labs' Medical LLM delivers **6, 7, and 8 points better average accuracy** than GPT-5, Gemini-2.5-Pro and Sonnet-4.5 respectively on MedHELM benchmarks

… with a model that **runs privately inside your environment on one GPU.**

Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation

Veysel Kocaman[1]; Mustafa Aytuğ Kaya[2]; Andrei Marian Feier[1]; David Talby[1]

Performance Benchmarks



nature › scientific reports › articles › article

Article | Open access | Published: 26 July 2025

Cancer type, stage and prognosis assessment from pathology reports using LLMs

Rachit Saluja, Jacob Rosenthal, Annika Windon, Yoav Artzi, David J. Pisapia, Benjamin L. Liechty & Mert R. Sabuncu

Scientific Reports **15**, Article number: 27300 (2025) | Cite this article

2268 Accesses | **2** Altmetric | Metrics

PathLLama
(State of the Art as of July 2025)

76.36%

Average Score

JSL-MedOnco
(November 2025)

90.73%

Average Score

Accuracy: Towards Regulatory-Grade Registry Data Abstraction

STUDY PARAMETERS

Dataset Volume
N=10,000 pathology reports processed.

Gold Standard
CTR validation (κ=0.92 agreement).

Noise Reduction

>95%

Non-reportable docs filtered pre-abstraction.

Time / Case

<2 min 120 min

98% reduction vs manual abstraction.

SYSTEM PERFORMANCE VS. BASELINE LLMS

Confidence Interval: 95%

Model / Approach	Primary Site Accuracy	Histology Accuracy	TNM Staging Accuracy	Hallucination Rate
Llama 3.1 8B Zero-shot, Open Source	68.4% ±2.1	62.1% ±2.3	45.8% ±3.1	18.2%
GPT-4o Few-shot, General Purpose	88.2% ±1.4	85.4% ±1.6	76.1% ±2.0	4.5%
Claude 3.5 Sonnet Few-shot, Long Context	91.5% ±1.1	89.2% ±1.3	81.3% ±1.8	2.1%
Hybrid System: John Snow Labs Medical SLMs + LLM + Rules	**98.4% ±0.4**	**97.6% ±0.5**	**94.2% ±0.8**	**<0.1%**

* $p < 0.001$ for all Hybrid System metrics vs baselines (McNemar's test)

Metric: Exact Match Accuracy vs CTR Ground Truth

Regulatory-Grade Governance

Redefining Real-World Evidence: John Snow Labs Introduces First FDA-Ready Patient Journey Platform

Present both supporting and contradictory evidence for every medical reasoning decision, to enable human review & audits

Highlight **where in each document** each fact was derived from

Store **full versioning & metadata** of which models, pipelines, and terminologies were applied

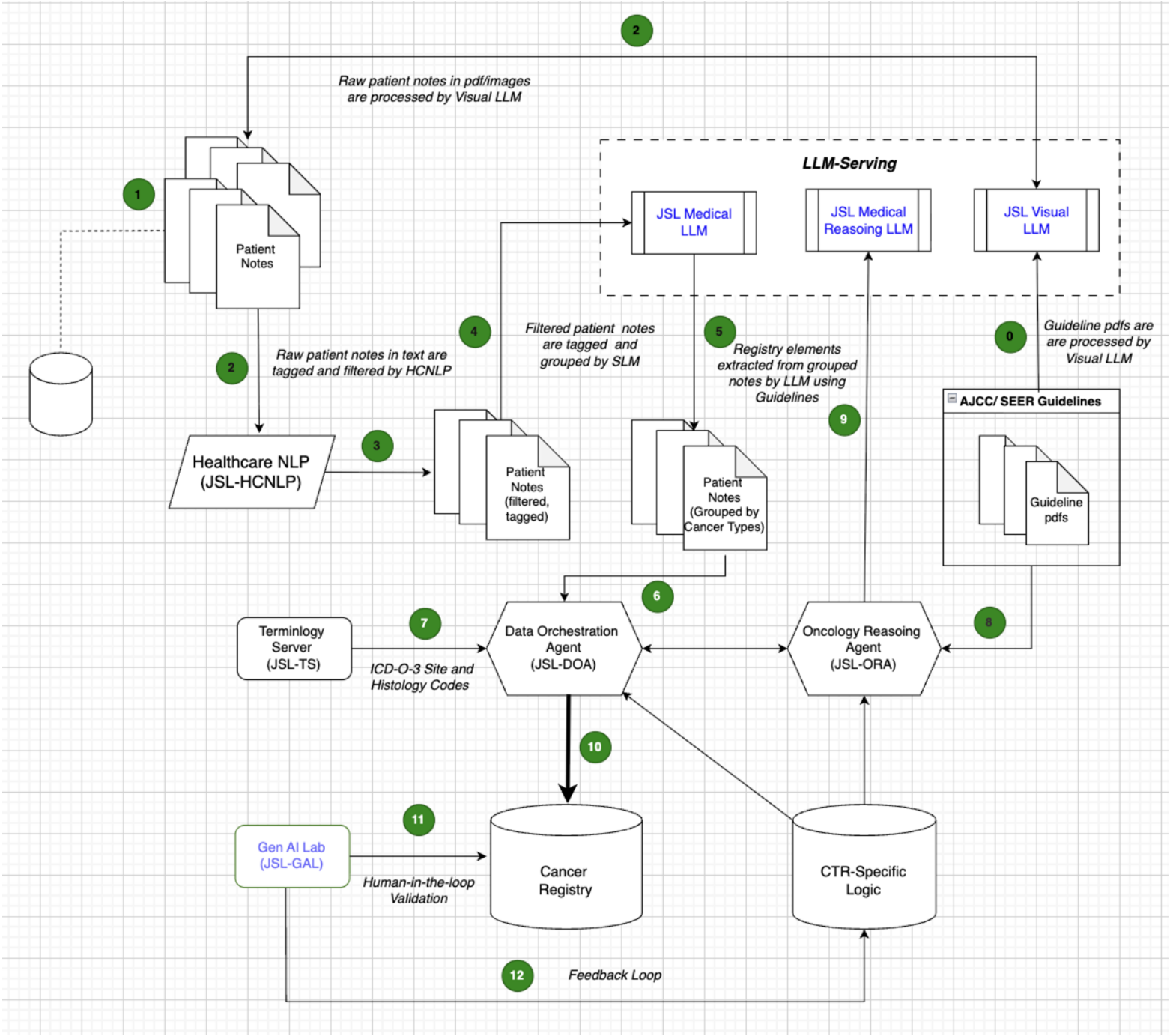
Solution Summary

Build the complete patient journey by continuously integrating multimodal data sources and grouping encounters to establish chronological context

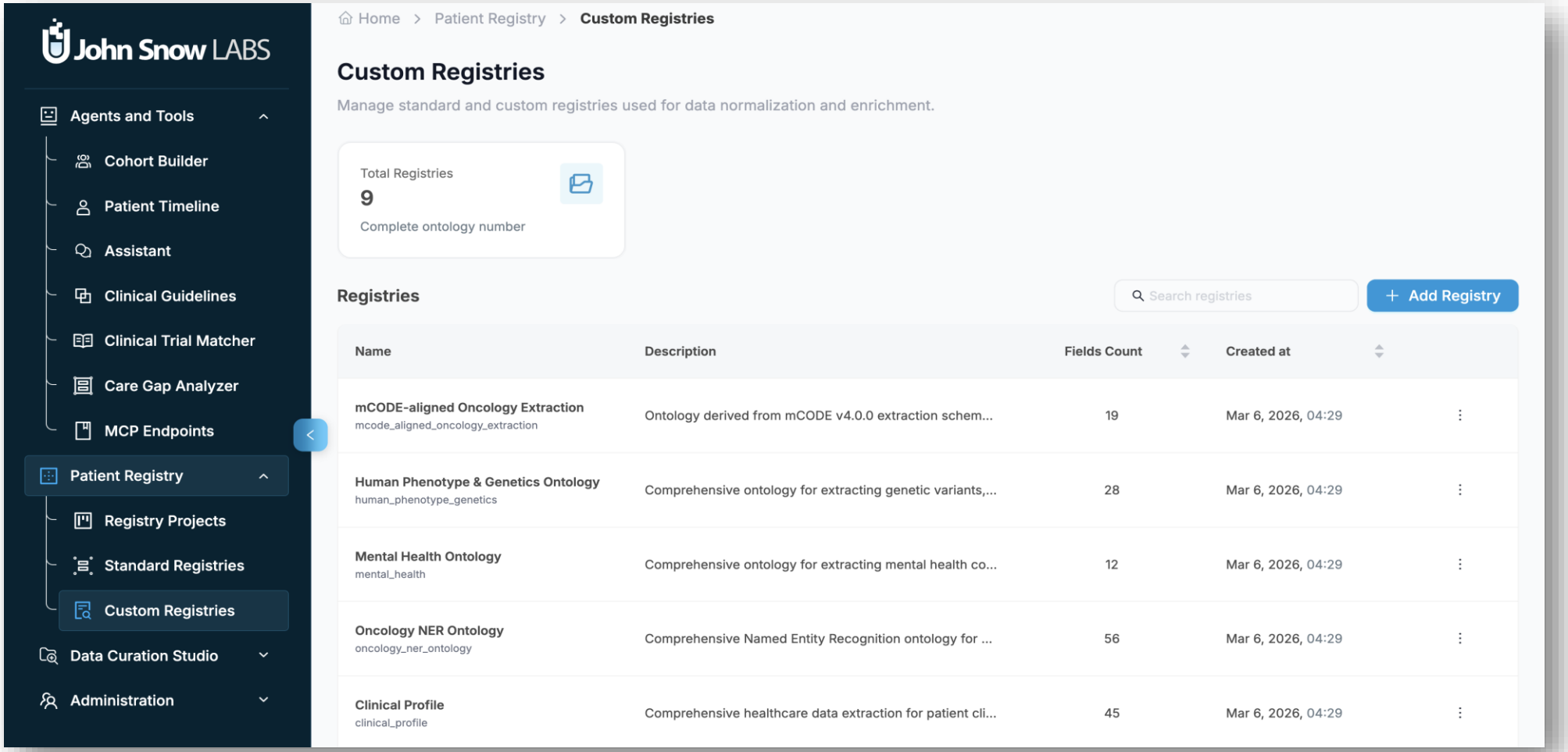
Combine healthcare-specific SLP, LLM, VLLM, and NLP models to handle thousands of pages per patient efficiently

Build narrow, role-specific agents to reduce context and raise accuracy when implementing medical reasoning

Regulatory-grade accuracy and governance are both needed for real-world, end-to-end registry automation



Ongoing Work



Built-in AI agents and tools for clinical guidelines, trial matching, care gaps, and cohort building

Define **new registries without coding**, benefitting from immediate end-to-end automation

Blind evaluation of accuracy versus certified registrars on a broader set of cancers and disease journeys

Learn More

40+ papers: **johnsnowlabs.com/peer-reviewed-papers**

80+ case studies: **johnsnowlabs.com/customers**

A Real-time NLP-Based Clinical **Decision Support Platform for Psychiatry and Oncology**

Applying **Healthcare-Specific LLMs** to Build **Oncology Patient Timelines** and Recommend Clinical Guidelines

AI-Enhanced Oncology Data: Unlocking Insights from EHRs with NLP and LLMs

Leveraging Healthcare NLP Models in **Regulatory Grade Oncology Data Curation**

Using **Healthcare-Specific LLMs** for Data Discovery from Patient Notes & Stories

Using **Generative AI for Data Extraction** Clinical Support

Large Language Models to Facilitate Building of **Cancer Data Registries**

Extracting what, when, why, and how from Radiology Reports in Real World Data Projects

Thank you!