



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?

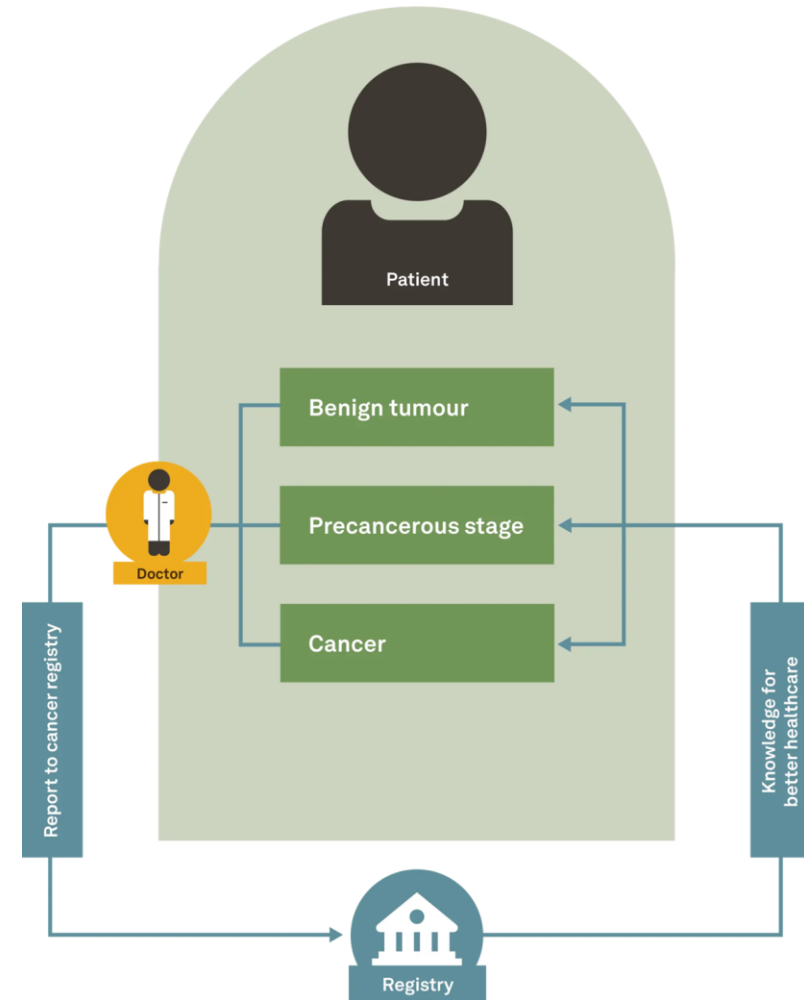
A systematic collection of data about cancer patients, including demographics, diagnosis details, treatments, and outcomes. These registries serve as **critical resources for monitoring cancer patterns** and improving patient care.

📌 Why Are They Needed?

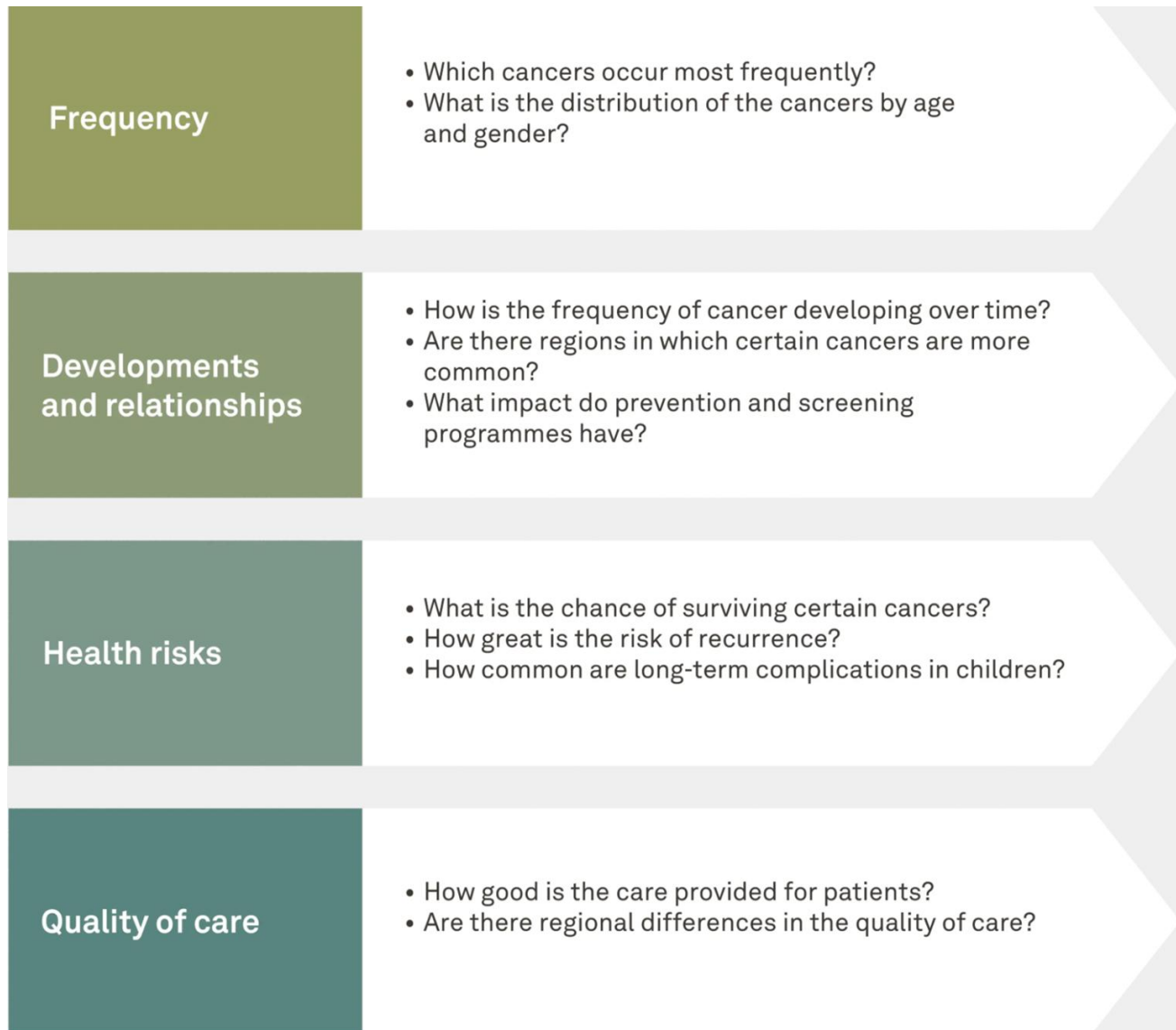
Cancer registries enable monitoring of incidence rates, survival trends, and treatment effectiveness. They inform **public health policies**, support clinical research, ensure quality of care, and facilitate international collaboration.

☰ 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

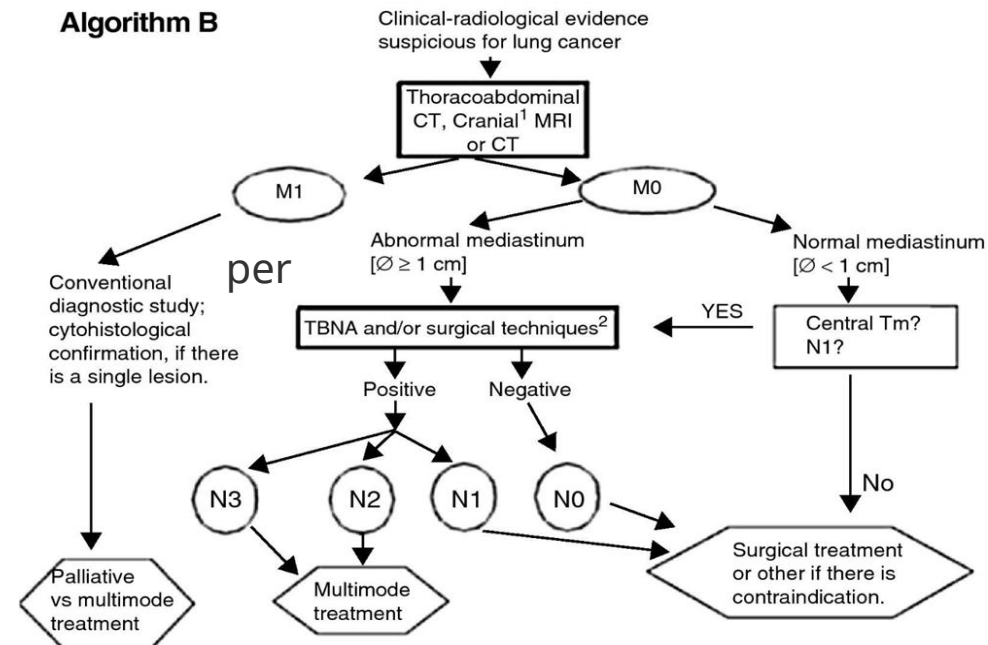
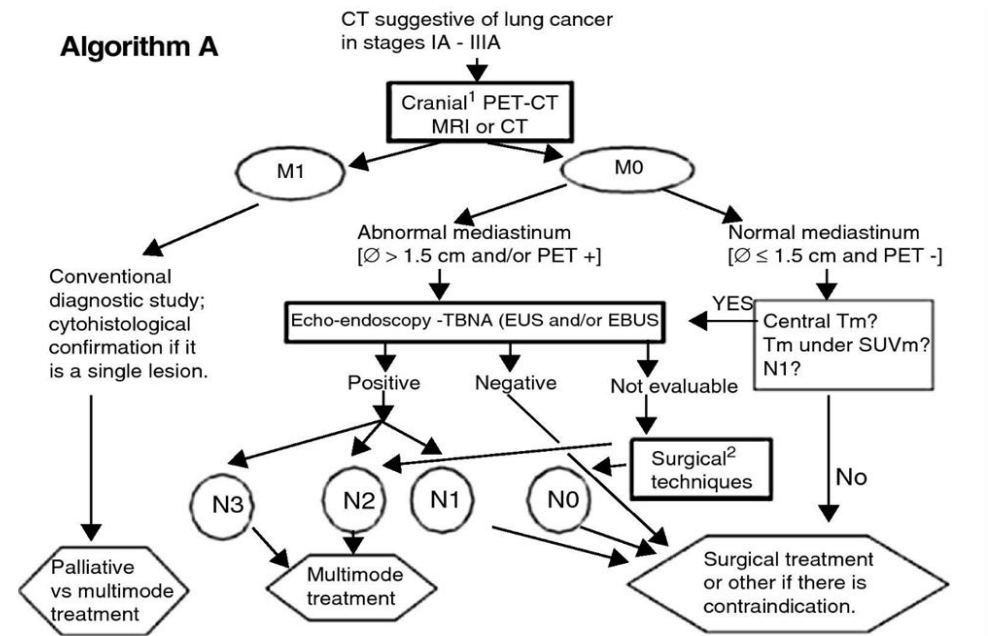
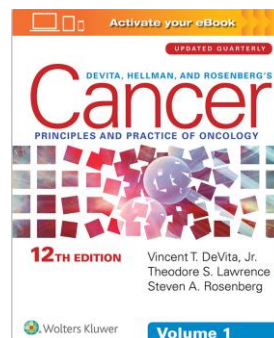
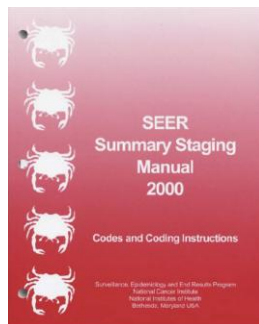
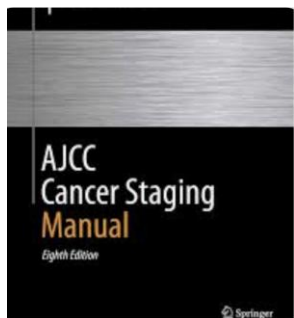
 Surgery Reports

 Genomic Tests

 Treatment Records

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

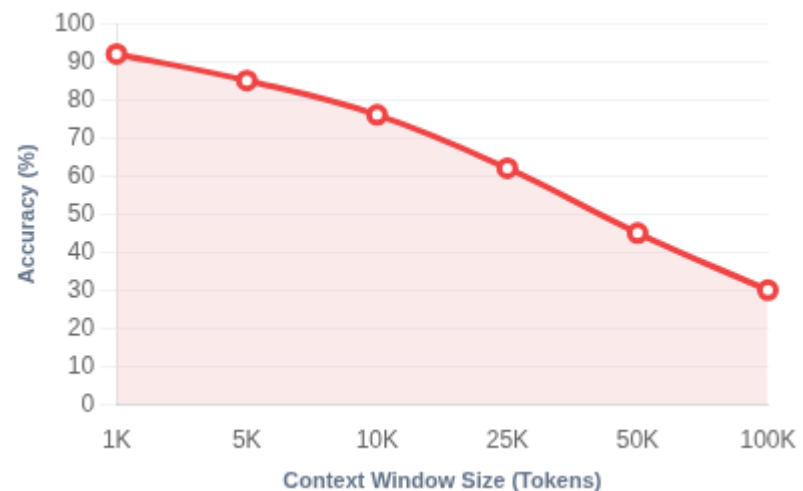
Workflows that registrars and auditors can use along with features for team collaboration, audit trails, and a learning loop that improves models over time.

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.

Deterministic Inference



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

Multimodal Data



📄 5,000+ Pages per Patient

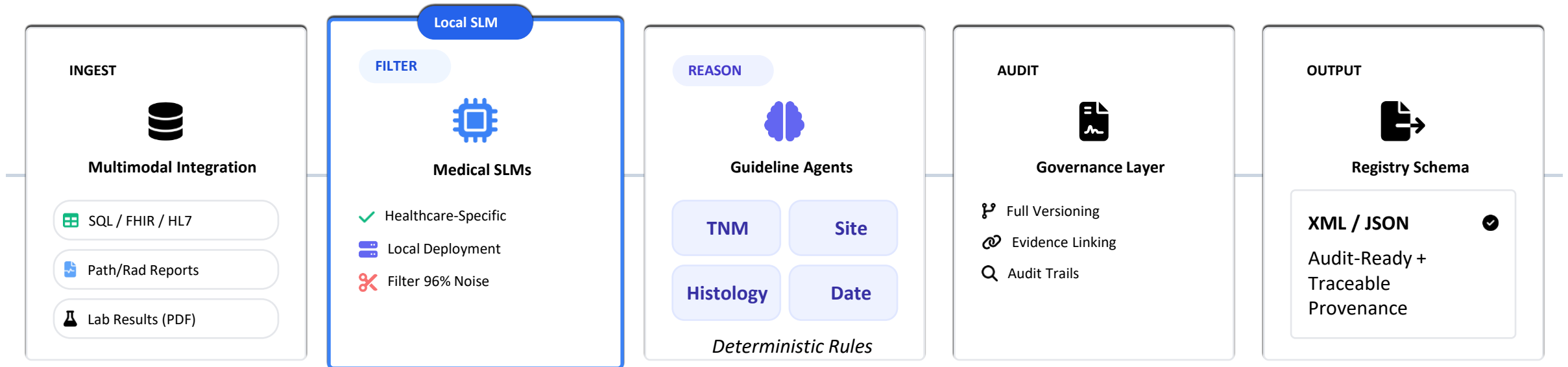
- × Requires integration of pathology (text), radiology (reports), and labs (tables).
- × Single-modality models miss cross-referencing cues essential for accurate staging.

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

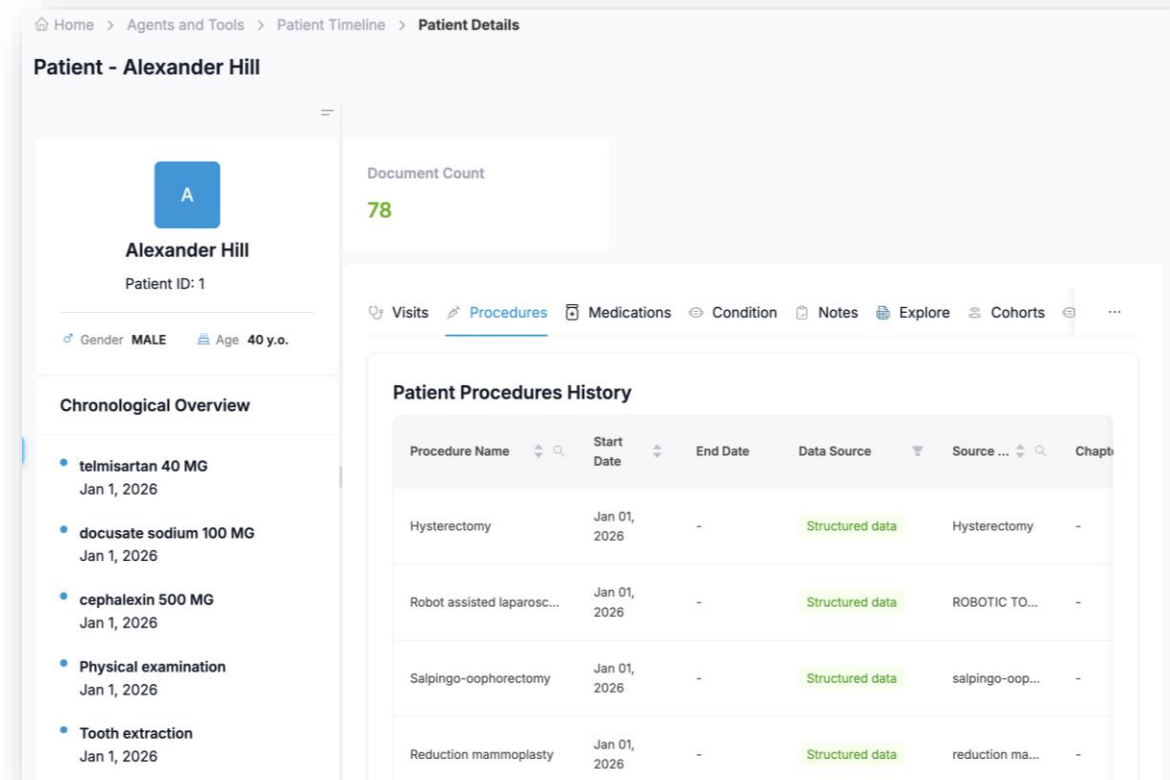
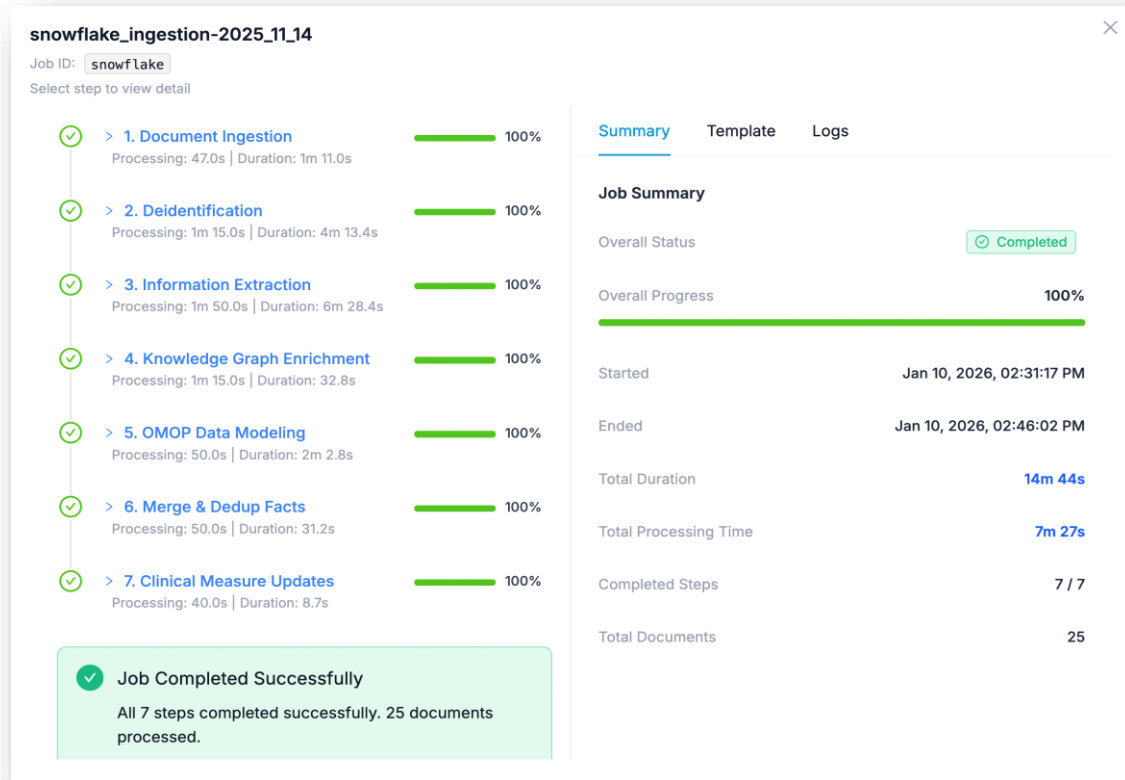
Bypasses high operational costs and privacy risks of frontier LLM APIs.



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

Home > Cohort Builder > New Cohort

Create Cohort

Build, analyze & compare patient cohorts with no-code rules and reusable filters.

Criteria Selection | Patient Management | Cohort Details | Review Cohort

Patient List

Name	Visits	Documents	Age
Mary Kim 4	18	28	22
James Allen 10	1	2	64
Crystal Martinez 14	4	12	46
Matthew Torres 28	1	3	79
Travis Shields 31	13	27	79
Terry Peterson 32	1	1	48
Joshua Dalton 33	4	4	60

Showing 1-10 of 273 items

Overview

Total Patients: 5458
Filtered Patients: 273

Advanced Filtering

Filter Group 1

Diagnosis OR

In

- Chronic obstructiv...
- Chronic obstructiv...
- Chronic obstructiv...
- Chronic obstructiv...

Out

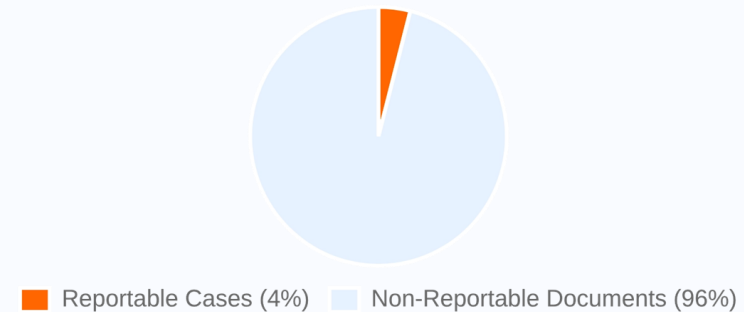
Select diagnoses...

+ Add Condition

Signal-to-Noise Ratio

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

ePath Report Reportability



- 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

Home > Data Curation Studio > **New Automation**

New Automation

1 Ontology Configuratio 2 Dataset Selection 3 Curation Info 4 Review

* Ontology

mCODE-aligned Oncology Extraction (Preset)

Don't see the ontology you need? [Manage Ontologies](#)

SELECTED ONTOLOGY

mCODE-aligned Oncology Extraction

Preset

Ontology derived from mCODE v4.0.0 extraction schema. Defines fields for document provenance, patient demographics, health assessment, primary/secondary cancer conditions, staging, biomarkers, treatment, disease status, and outcomes. Each field includes extraction guidance and example values.

FIELDS CONFIGURATION

Field	Type	Category	Required
File Name / Document Identifier file_name	text	document	Yes
Patient Identifier patient.identifier	text	IDENTIFIER	Yes
Patient Name patient.name	text	demographics	No
Date of Birth patient.birthDate	date	demographics	No
Gender patient.gender	text	demographics	No



Home > Data Curation Studio > **Curation Details**

Review & Validation - Cancer Registry

Original Export

Selected Patients (1)

Jeffrey Chang
62 yrs • MALE
76 docs

Overview Staging Raw Result Version History

Cancers Summary

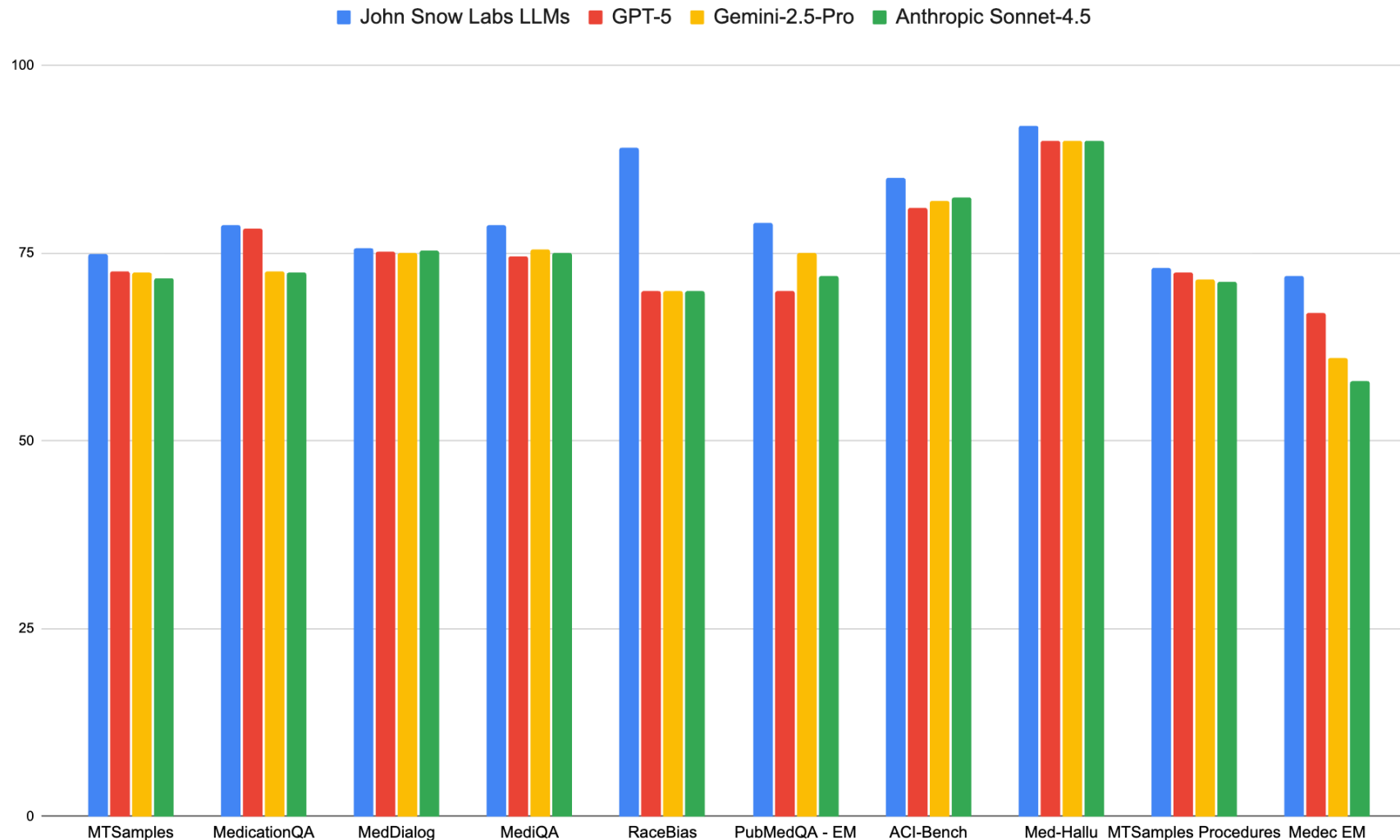
	Primary Site	Histology	Diagnosis Date
+	Appendix (C18.1)/ Malignant C18.1	8480/3 - Mucinous adenocarcinoma 8480/3	2005-05-17
+	Prostate (C61)/ Malignant C61	8140/3 - Adenocarcinoma, NOS 8140/3	2015-07-26
+	Bladder, posterior wall (C67.4)/Malignant C67.4	8120/3 - Transitional cell carcinoma, NOS 8120/3	2019-06-07

Cancer Registry Overview

Total Cancers	3	Registry Ready	3 / 3
	pt39_doc71	pt39_doc3	pt39_doc6
	pt39_doc20	pt39_doc21	pt39_doc27

- 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



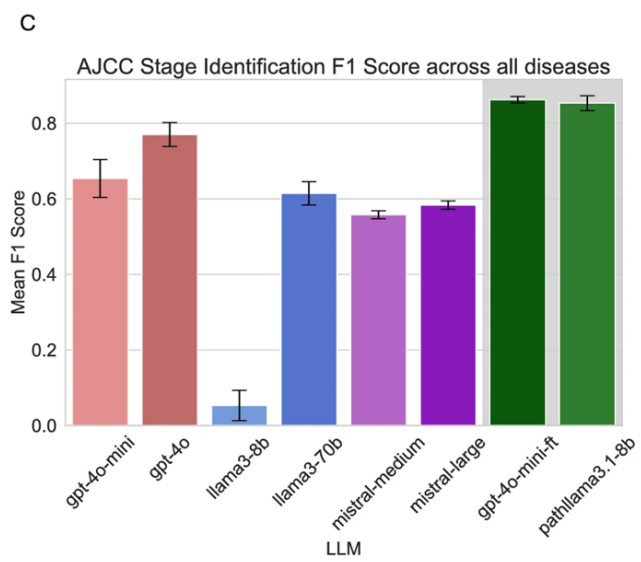
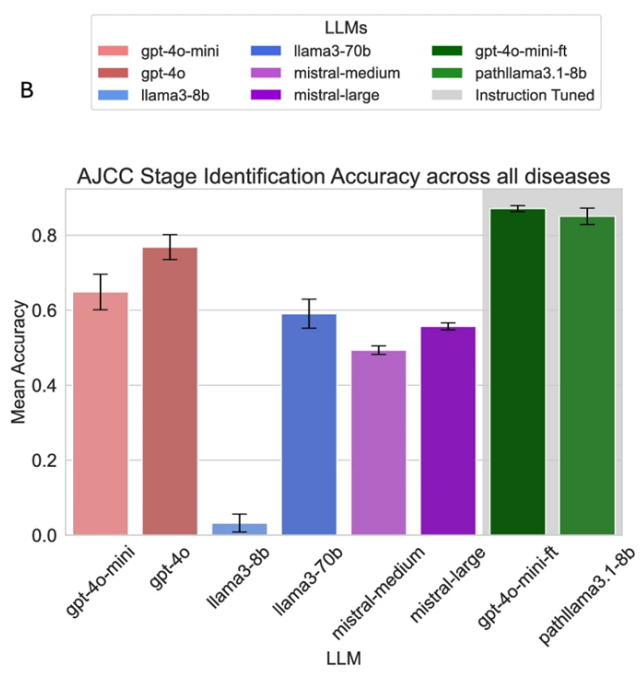
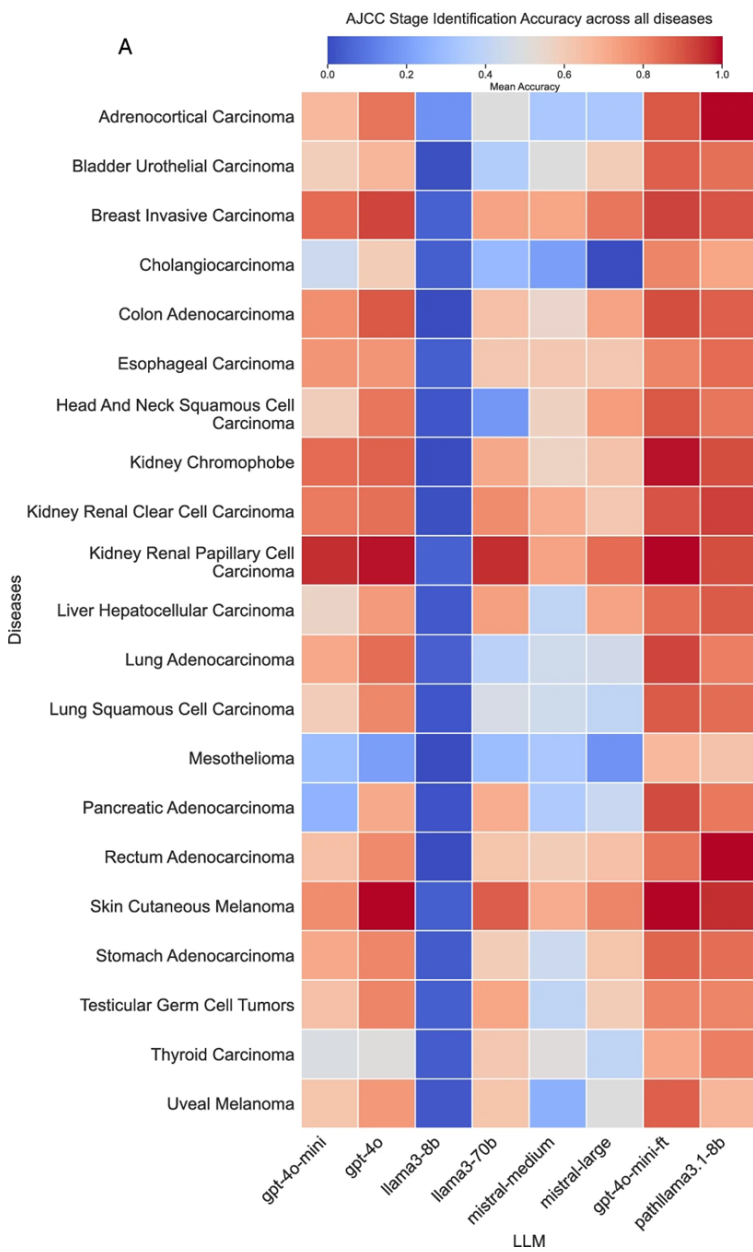
John Snow Labs' Medical LLM delivers **6, 7, and 8 points better average accuracy** than GPT-5, Gemini-2.5-Pro and Sonnet-4.5 respectively on MedHELM benchmarks

... 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 Kocaman¹](#); [Mustafa Aytuğ Kaya²](#); [Andrei Marian Feier¹](#); [David Talby¹](#)

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 ($\kappa=0.92$ 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

Evidence: cancer_1 > Diagnosis Date

2002-07-07

Evidences

5 evidence items were found (2 supporting, 3 contradictory), from 4 different documents.

- Supporting Evidence(2)

pt47-doc1 2002-07-07 95%

Clinical note documents the original diagnostic prostate biopsy date of 07/07/2002 corroborating pathology biopsy collection — earliest pathologic confirmation chosen as NAACCR diagnosis date.

pt47-doc75 2002-08-26 80%

Biopsy specimen collection noted with 07/07/02 in the pathology specimen header; supports earliest pathologic collection date selection.

- Contradictory Evidence(3)

pt47-doc5 2002-09-07 95%

Multiple clinic/pathology notes reference the prostatectomy date 09/07/2002 as a pathologic event; these are later pathologic dates and therefore contradictory when considering the earliest biopsy date as diagnosis date.

Status:
Rising PSA, non castrate.

Visit Details:
The history was obtained from the patient and an outside medical record. The patient arrived walking .

Preferred Language:
The patient states their preferred language for health care discussion is English .

History of Present Illness:
Mr. William is a 72-year-old man who was originally diagnosed with prostate adenocarcinoma in 2003. Prostate biopsy on 07/07/2002 identified Gleason 3+3 disease. He believes his PSA was in the 6-7 NG/ML range at that time.

He underwent radical prostatectomy in 09/07/2002 which identified Gleason 3+4 disease and no evidence of extracapsular extension or seminal vesicle invasion. Surgical margins negative. 0/8 lymph nodes involved. Final staging pT2bN0.

After surgery his PSA was undetectable however later increased and he received salvage radiation therapy between Apr. 2006 and Jun. 2006. His PSA was 0.08 at the time of radiation and then nadired again to undetectable levels for several years.


By 2012 his PSA was 0.11 and by Jan. 2015 his PSA had increased to 0.5. He then started on a clinical trial using metformin in Jan. 2015. He was on this trial for fixed period of time -he believes about 8 months. His PSA was undetectable on two occasions during that trial and then subsequently increased from 0.2 to 1.1 NG/ML. He came off study. On 04/03/16 his PSA was 1.5 NG/ML.

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

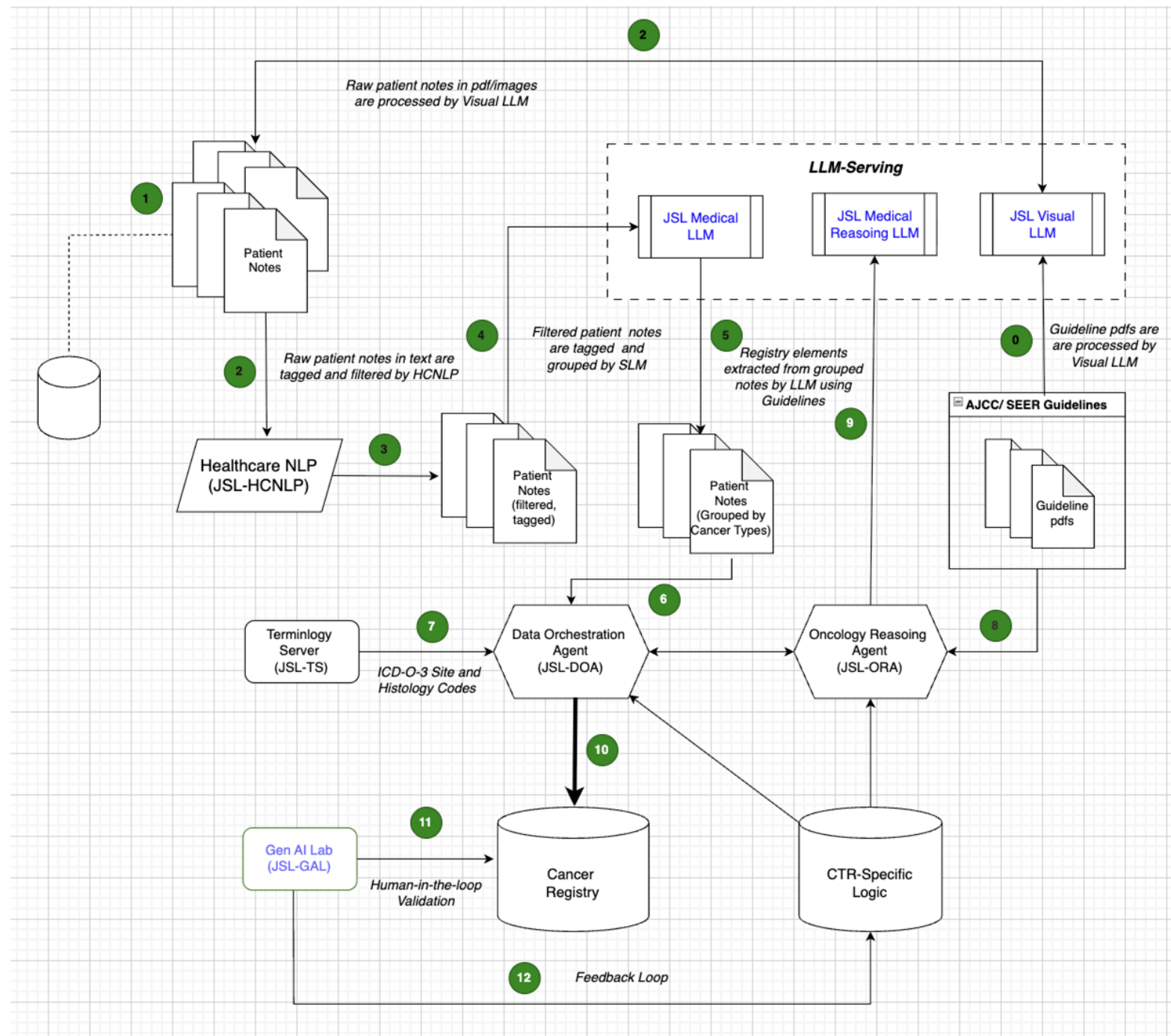
Solution Summary

 **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

John Snow LABS

Home > Patient Registry > Custom Registries

Custom Registries

Manage standard and custom registries used for data normalization and enrichment.

Total Registries **9**
Complete ontology number

Registries

Search registries + Add Registry

Name	Description	Fields Count	Created at	
mCODE-aligned Oncology Extraction mcode_aligned_oncology_extraction	Ontology derived from mCODE v4.0.0 extraction schem...	19	Mar 6, 2026, 04:29	⋮
Human Phenotype & Genetics Ontology human_phenotype_genetics	Comprehensive ontology for extracting genetic variants,...	28	Mar 6, 2026, 04:29	⋮
Mental Health Ontology mental_health	Comprehensive ontology for extracting mental health co...	12	Mar 6, 2026, 04:29	⋮
Oncology NER Ontology oncology_ner_ontology	Comprehensive Named Entity Recognition ontology for ...	56	Mar 6, 2026, 04:29	⋮
Clinical Profile clinical_profile	Comprehensive healthcare data extraction for patient cli...	45	Mar 6, 2026, 04:29	⋮

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

COTA

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!