WHITE PAPER

John Snow LABS

Why Secondary Use of Clinical Data Demands Multimodal AI

Up to 87% of the clinical information that defines a patient's journey never appears in the structured fields your analytics teams depend on.
The consequences are measurable.

40%

of diagnoses are missed by structured data

46×

more SDOH found in clinical notes vs. ICD-10 codes

87%

of clinical facts exist only in free-text notes

68%+

of cancer staging absent from structured data

Published by John Snow Labs — Patient Journey Intelligence

www.johnsnowlabs.com/pji/

Contents

01. Quantify the Data Accuracy Gap

Multiple independent, peer-reviewed studies find that across every clinically meaningful data type, the gap between what is documented and what is captured in structured fields is not marginal, it is catastrophic.

Data Type	Accuracy Gap	Clinical & Business Implication
Diagnoses	**40% missed** (Poulos et al. 2021)	Half your eligible patients are invisible to cohort queries
Family History	**12x gap** (Polubriaginof et al. 2015)	Risk stratification models lose their most heritable signal
Social Determinants	**46x gap** (Guevara et al. 2024)	Readmission models ignore the social drivers of most variation
Cancer Staging	**68%+ missing** (Emamekhoo et al. 2022)	Registry and RWE studies cannot accurately report stage distribution
Medication Histories	**>60% errors** (Lombardi et al. 2016)	Pharmacovigilance studies built on a structurally wrong foundation
Suicide / Self-Harm	**>81% missed** (Fernandes et al. 2018)	Structured risk models blind to overwhelming majority of at-risk patients
All Clinical Concepts	**87% in text** (Seinen et al. 2025)	AI on structured fields operates on <1/7 of available clinical signal

02. Executive Summary

Every healthcare organization running analytics, building AI models, or operating patient registries believes it is working with clinical data. The uncomfortable truth: most are working with a carefully curated fragment of that data, the part that was easy to collect, not the part that is clinically meaningful.

Decades of peer-reviewed research are unambiguous: structured EHR fields or the codes, tables, and checkboxes that feed the vast majority of analytics pipelines, capture as little as 13% of the clinical information documented during patient care. The remaining 87% lives in physician notes, discharge summaries, pathology reports, and operative records: the narrative layer where clinicians actually communicate what happened to their patients.

This is not a data quality problem that can be patched with data governance policies. It is a structural incompleteness problem and it is silently corrupting the cohorts, models, registries, and real-world evidence studies that organizations are betting their research and AI strategies on.

This white paper:

- Quantifies the accuracy gap across critical clinical data types, backed by peer-reviewed studies
- Explains why structured-only analytics produce systematically wrong, not just incomplete, results
- Defines the eight capabilities a modern secondary-use data platform must deliver
- Shows how patient-level reasoning resolves conflicting, temporal, and cross-document clinical evidence
- Explains how AI agents built on complete patient data unlock new operational workflows
- Explains how Patient Journey Intelligence closes this gap without bespoke pipeline development

03. The Dashboard Delusion: How "Invisible" Cohort Attrition Corrupts Research and AI Strategy

Ask any Chief Data Officer or VP of Clinical Informatics whether they trust their analytics data, and most will say yes. They have data warehouses, FHIR endpoints, curated EHR feeds, and dashboards. What they rarely have is a clear picture of how much clinical reality is missing from those systems.

The problem is not visible in the dashboards. A cohort query that misses 40% of eligible patients still returns a number. A risk model trained on incomplete data still produces a score. A cancer registry built on structured fields alone still generates reports. The gap is silent, until the results start diverging from clinical reality, trials fail to enroll, or regulators start asking for provenance.

And that gap is widening. As AI scribes become more common, more of the clinically meaningful story is being captured in narrative form. These tools may improve note completeness and reduce physician burden, but they also increase the volume and richness of unstructured documentation. If organizations continue to rely primarily on structured fields for secondary use, a growing share of patient truth will sit outside the datasets that power research, quality measurement, and AI.

The Structured-Only Trap

Structured EHR data was designed for billing, documentation, and care coordination, not for complete clinical representation. When organizations rely on it exclusively for research or secondary use, they are not making a data quality decision. They are making a data completeness decision that systematically distorts every downstream output.

AI scribes intensify this problem. They capture nuance, context, and longitudinal detail more effectively than traditional documentation workflows, but that value often remains locked in free text. The result is a widening divide between what clinicians know, what the chart says, and what analytics systems can actually measure.

04. The Cost of Fragmented Clinical Evidence

The consequences of operating on incomplete clinical data are not abstract. They manifest in specific, measurable business and clinical failures:

Clinical Trial Recruitment Falls Short

Cohort queries that miss 40% of eligible patients don't just slow enrollment, they structurally under power studies. Teams compensate by extending timelines and expanding sites, neither of which addresses the root cause.

AI Models That Cannot Scale

Models trained on structured-only data operate on less than one-seventh of the available clinical signal. They can achieve high accuracy on narrow tasks with clean labels, but fail to generalize. The models are not the problem. The training data is.

Registries That Are Costly and Slow to Maintain

Cancer staging is missing from structured EHR fields in more than 68% of encounters (Emamekhoo et al. 2022). That means the majority of staging data must be manually abstracted by trained staff reviewing pathology reports and clinical notes one record at a time. Manual abstraction is expensive, time-consuming, and difficult to scale - the bottleneck that delays registry reporting cycles, constrains research cohort sizes, and limits the frequency with which registry data can be refreshed. Replacing manual chart review with automated AI extraction is not just a speed advantage; it is the only viable path to registries that keep pace with the volume of incoming clinical data.

05. The Unstructured Layer: Overcoming the Technical Debt of Structured-Only Clinical Architectures

The data accuracy gap is not a consequence of poor documentation. Clinicians write extensive, rich narratives. The problem is that the infrastructure built to repurpose this data for secondary use systematically ignores the most information-dense layer: unstructured clinical text.

This happens for understandable reasons. Extracting clinical facts from free text at scale requires healthcare-specific NLP, medical ontology mapping, temporal reasoning, and clinical conflict resolution, capabilities that are expensive and complex to build. Organizations default to structured data not because it is better, but because it is accessible.

The result is a self-reinforcing cycle: because structured data is what teams use, infrastructure evolves to serve structured data better, making the unstructured layer even more inaccessible, widening the accuracy gap further with each passing year.

Which one do you want your AI to learn from?

diabetes mellitus

Disease_Syndrome_Disorder

E11

Type 2 diabetes mellitus

Over the last year, this 58-year-old[Age] male[Gender] has had poorly controlled[Modifier] diabetes mellitus (Disease_Syndrome_Disorder: E11 Type 2 diabetes mellitus) with an HbA1c (Test: 86910-7 Hemoglobin A1c/Hemoglobin.total goal Blood) of 11.2[Test_Result] %, following maximum metformin (Drug_BrandName: 6809) dose. He has had three recurrent[Modifier] DKA (Disease_Syndrome_Disorder: E11 Type 1 diabetes mellitus with ketoacidosis) admissions and has been non-adherent[Modifier] to his insulin (Drug_BrandName: 67866001) regimen despite repeated counseling.

The patient has also developed diabetic neuropathy (Disease_Syndrome_Disorder: E1140 Type 2 diabetes mellitus with diabetic neuropathy, unsp) .

Sterile EHR - ICD-10 Code

Rich, predictive narrative with coded entities

Why Text Outperforms Codes

A patient coded with "diabetes" tells you nothing about glycemic control, medication adherence, complications, or trajectory. A clinical note documents: "HbA1c 11.2%, poorly controlled despite max-dose metformin, recurrent DKA admissions, non-adherent to insulin." The predictive signals that actually drive outcomes live in that narrative and remain invisible to structured-only analytics.

06. The Secondary Use Checklist: Non-Negotiable Capabilities for Multimodal Clinical Data Integrity

Closing the clinical data accuracy gap requires more than better tooling. It requires a different architectural philosophy, one built around multimodal completeness rather than structured convenience. Here are the eight non-negotiable capabilities:

01 | Multimodal Clinical Data Integration

Ingest across all modalities: structured EHR, clinical notes, scanned PDFs, imaging metadata, FHIR resources, labs, claims. Relying on any single modality by design excludes critical information.

02 | Healthcare-Specific NLP

Extract diagnoses, medications, findings, procedures, social determinants of health, from free text using medical language models that understand negation, uncertainty, and assertion status, with over 85–95% accuracy.

03 | Terminology Standardization

Map all concepts to SNOMED CT, RxNorm, LOINC, and ICD-10-CM. Without normalization, the same clinical fact appears in dozens of forms, making accurate patient counts and cross-institution research impossible.

04 | Clinical Reasoning & Conflict Resolution

Resolve conflicts, deduplicate entities, and distinguish confirmed diagnoses from ruled-out conditions. Real-world data is messy, without intelligent reasoning, systems amplify noise rather than surface signal.

05 | Longitudinal Patient Timelines

Organize all clinical events chronologically with precise temporal context. Most clinical questions involve time: disease progression, treatment response, time-to-event outcomes cannot be answered without it.

06 | Privacy & De-Identification

Remove PHI automatically using HIPAA and GDPR-compliant methods at 99%+ accuracy. Manual approaches are slow, expensive, and error-prone and specialized medical patterns require specialized tools.

07 | Provenance & Auditability

Track complete lineage from source document to final output. Regulatory compliance, research reproducibility, and clinical trust all require transparency, not black boxes.

08 | Continuous Updates for Living Datasets

Keep patient journeys continuously updated as new data arrives, not static snapshots that become stale. Clinical data accumulates continuously; quarterly refreshes guarantee obsolete results.

07. Patient-Level Reasoning: Beyond Simple Extraction

Extracting clinical facts from individual documents is a necessary first step - but it is not sufficient. A patient's true clinical picture is distributed across dozens of encounters, written by different clinicians, at different points in time, using different terminology, and sometimes reaching contradictory conclusions about the same condition.

The challenge of secondary use is not just reading documents. It is synthesizing evidence across documents into a coherent, temporally accurate, clinically trustworthy representation of what actually happened to a specific patient.

Four Dimensions of Patient-Level Reasoning:

01 | Conflict resolution

When one encounter rules out a diagnosis that another confirmed, the system must determine which assertion is authoritative, based on date, source type, and clinical context.

02 | Temporal reasoning

Diagnoses evolve. 'Suspected pulmonary embolism' at admission becomes 'confirmed PE' post-imaging, then 'resolved' at discharge. The timeline must reflect this progression, not collapse it into a single, undated flag.

03 | Assertion detection

Clinical text is full of negations, hypothetical, and family-history references. 'Patient denies chest pain,' 'rule out MI,' and 'mother had breast cancer' must be correctly attributed, not recorded as findings of the patient.

04 | Deduplication and source merging

A medication appearing in a discharge note, a structured medication list, and a scanned referral letter represents one drug, not three. Merging these records requires source-confidence scoring and reasoning, not simple string matching.

Scenario	Naive Approach	Patient-Level Reasoning
Conflicting diagnoses across notes - COPD documented in one encounter, ruled out in another	Treats both records as confirmed, creating duplicate or contradictory disease flags	Assertion detection resolves conflict; only the affirmed diagnosis is retained in the structured timeline
Diagnosis that changes over time - “suspected cancer” at admission, confirmed post-pathology	Stages the patient with the provisional label, distorting registry and RWE staging data	Temporal reasoning tracks the diagnosis lifecycle; the confirmed post-pathology status is the authoritative value
Medication appears in three sources: discharge note, medication list, and a scanned referral	If all three records are treated as independent, medication is counted three times or silently dropped	Deduplication and source-confidence scoring merge records into a single, attributed medication entry
Family history mentioned in oncology note, not in structured family history fields	Structured-only query returns no family history and the model misclassifies patient’s risk tier	NLP assertion extraction flags family-context assertion; entity is assigned to relatives, not the patient

Without patient-level reasoning, structured timelines contain ghosts diagnoses that were never confirmed, medications counted multiple times, risk factors attributed to the wrong person. The accuracy gap is not just about what is missing from structured fields. It is also about what is misrepresented when unstructured data is extracted without the clinical intelligence to interpret it correctly.

08. Agentic AI on Complete Patient Data

A complete, continuously updated patient timeline is not just better data for existing analytics, it is the foundation that makes a new class of AI application possible: clinical agents. But deploying agents that healthcare and research teams can actually trust places demanding requirements on the underlying data platform.

Clinical agents are AI systems that can autonomously execute multi-step tasks against a patient data foundation: identifying cohorts matching complex criteria, retrieving supporting evidence for a clinical question, or monitoring incoming data for quality issues, without requiring a human to manually query records or write bespoke extraction logic for each new question.

The critical distinction between agents that work and agents that fail is the quality of the data they operate on. An agent querying a structured-only dataset inherits all of its incompleteness. An agent operating on a complete, multimodal, reasoning-enriched patient timeline can return answers that reflect clinical reality.

Agent Type	What It Does	Primary Use Cases
Cohort & Registry Agent	Dynamically identifies patient cohorts against complex, multi-criteria definitions by querying the full longitudinal timeline – structured and unstructured – not just coded fields. Automates registry abstraction tasks that previously required manual chart review.	**Primary use cases:** Clinical trial recruitment, disease registry population, quality measure reporting
Clinical Evidence Agent	Traverses patient timelines to surface supporting evidence for a clinical question: which patients had a specific progression event, which were on a specific drug class within a defined window, which have a confirmed biomarker result. Returns citations to source documents.	**Primary use cases:** Real-world evidence generation, comparative effectiveness, pharmacovigilance
Data Curation & QA Agent	Continuously monitors incoming data for gaps, conflicts, and quality issues across all modalities. Flags records that require resolution, applies confidence scoring, and keeps the shared patient timeline accurate as new clinical data arrives.	**Primary use cases:** Data quality governance, living dataset maintenance, regulatory audit preparation

The Data Foundation Is the Prerequisite

Clinical agents are only as reliable as the data they query. An agent built on a structured-only dataset will confidently return wrong answers, it cannot know what it cannot see. The eight capabilities outlined in Section 06, from multimodal ingestion through patient-level reasoning and continuous updates, are the non-negotiable prerequisites for deploying agents that clinical and research teams can actually trust.

09. FDA Requirements for Real-World Evidence: Data Accuracy and Full Provenance

The regulatory landscape for Real-World Evidence (RWE) has evolved rapidly. The U.S. Food and Drug Administration has issued clear and binding guidance on what is required for medical devices and regulatory decision-making using real-world data. For organizations building clinical AI for regulatory purposes, two expectations are non-negotiable.

01 | Capture Complete Clinical Information, Not Just Claims Data

Claims data alone is insufficient for regulatory-grade evidence. The FDA's guidance on the use of real-world evidence to support regulatory decision-making for medical devices explicitly identifies data quality and completeness as foundational requirements. Critical clinical signals such as diagnoses, outcomes, adverse events, and treatment response, are frequently missing or incomplete in structured fields.

This is not abstract guidance. Cohorts built from claims data will be rejected if critical clinical details exist in unstructured notes that were not captured. Registries missing 40% of diagnoses do not meet regulatory scrutiny. Safety surveillance systems that rely only on structured ICD codes will miss adverse events documented in clinical narratives.

02 | Trace Every Result Back to Its Source Documentation

When submitting evidence to regulators, aggregate numbers are not enough. Every result must be explainable, reproducible, and auditable. Regulators must be able to trace results back to: individual patients included in cohorts; source documents supporting each clinical fact; dates, forms, and clinical context; and the transformations and model versions used in processing.

A model that produces accurate predictions but cannot explain its inputs is insufficient. A cohort builder that identifies patients but cannot cite source documentation is insufficient. A registry that extracts data elements but cannot trace them to original clinical notes is insufficient.

Regulatory Expectations Are Increasing, Not Decreasing

The December 2025 final FDA guidance on real-world evidence for medical device regulatory decision-making signals that expectations for RWE are becoming more stringent. Organizations that cannot demonstrate data accuracy, provenance, and reproducibility will face increasing challenges in regulatory submissions.

Building these capabilities after the fact is far more expensive than designing them in from the start. Any platform used for regulatory-grade secondary use must treat FDA RWE requirements as design constraints. Complete lineage from source document to final output, model versioning, confidence scoring, and full auditability must be foundational platform capabilities, available to every application without bespoke engineering per submission.

10. Governance: What Every Secondary-Use Platform Must Implement

Clinical AI is not just a data engineering challenge, but also a governance challenge. Healthcare organizations that build AI on patient data must satisfy requirements that go far beyond analytics accuracy: HIPAA compliance, GDPR obligations, institutional data policies, and FDA expectations for real-world evidence. These requirements are not optional additions to a working system. They are foundational design constraints that must be built in from the start, not retrofitted after deployment.

Organizations that attempt to add governance capabilities to systems that were not designed for them consistently face the same outcome: the architecture does not support it without a rebuild. The six requirements below define what any secondary-use clinical data platform must implement to be deployable in production.

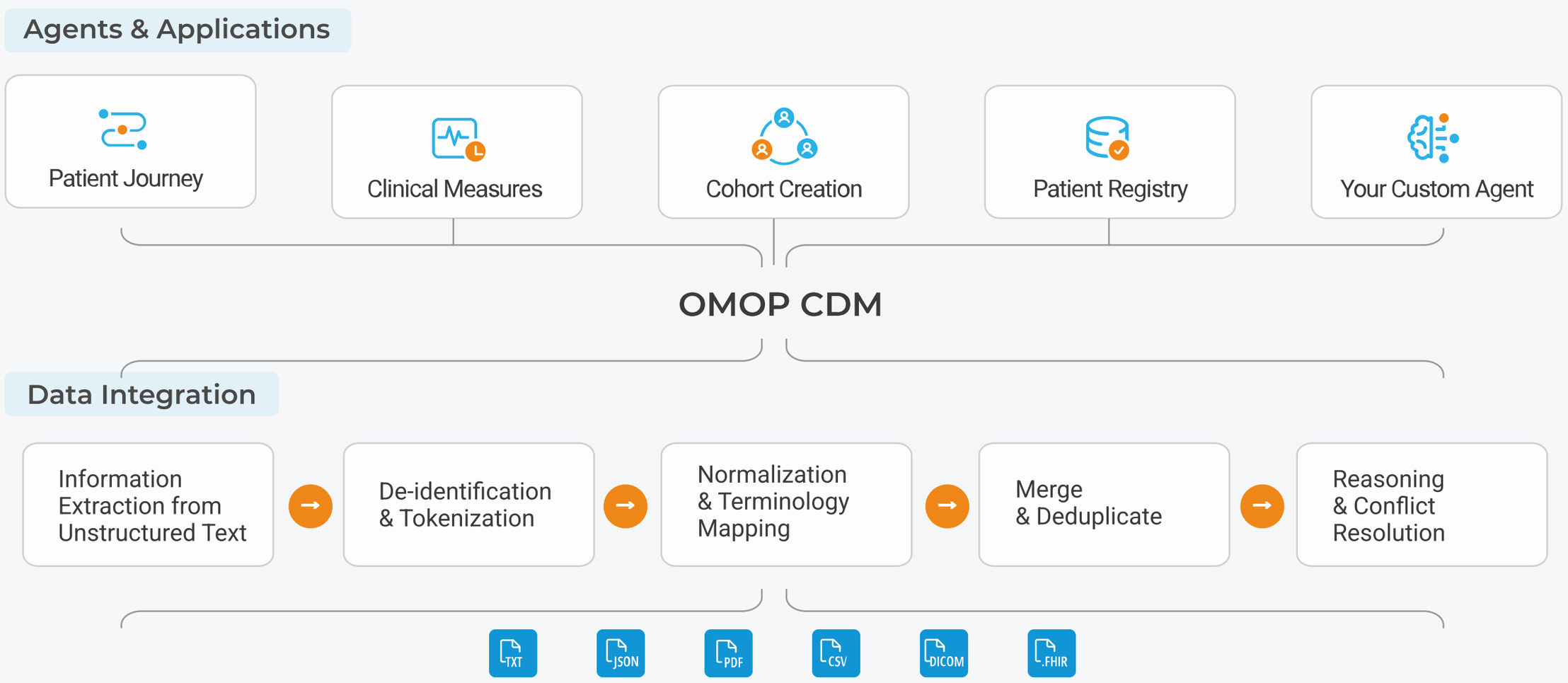
Requirement	Why It Cannot Be Optional
PHI De-Identification & Parallel Datasets	Most secondary-use workflows require de-identified data: research collaborations, AI model training, external submissions. But de-identified data separated from its identified counterpart becomes stale and inconsistent. A platform must maintain parallel identified and de-identified datasets from the same source, kept semantically synchronized, so operational and research teams never diverge on clinical facts.
Fine-Grained Role-Based Access Control	Different users, roles, and AI agents need different data permissions. A clinical trial coordinator should not access mental health or billing records. A registry abstractor should be scoped to their cancer site. An AI agent should be constrained to data it has been authorized to query. Access policies must be enforced at the data level, not just at the application level, so that bypassing the front end does not bypass the governance.
Comprehensive Audit Logging	Every data access, extraction event, tool invocation, and agent action must be logged with user identity, timestamp, and clinical context in tamper-evident logs retained per institutional policy. When a regulator, IRB, or legal team asks 'who accessed what, when, and for what purpose,' the answer must be complete and retrievable without manual reconstruction.
Provenance & Lineage for Every Derived Fact	Each clinical fact used in analytics, a registry, or an AI recommendation must carry a complete provenance chain: source document, document date, extraction method, model version, confidence score, and precise text location — preserved through every transformation including NLP extraction, terminology normalisation, and OMOP mapping. Without end-to-end lineage, regulatory submissions, research reproducibility, and clinical trust are all impossible.
Private, Secure Deployment	PHI must not flow to external services, shared cloud environments, or third-party model providers. Medical language models must run locally. Air-gapped deployments must be fully supported. This is a compliance requirement for most healthcare organizations, not a preference. A platform that cannot operate entirely within institutional infrastructure is not suitable for clinical AI production.
Versioning & Reproducibility	Results from any analysis must be exactly reproduced using the data and model versions that produced them. This supports regulatory reproducibility requirements (21 CFR Part 11, FDA RWE guidance), retrospective audit, and comparison of results across model update cycles. Without time-stamped versioning, any regulatory challenge to a historical result requires manual reconstruction that may be impossible.

11. Transitioning from Fragmented EHR Feeds to a Unified OMOP Foundation with Patient Journey Intelligence

The previous sections have built a case from peer-reviewed evidence. They have clearly established, that structured-only clinical data misses 40–87% of clinically relevant information. They have defined eight non-negotiable data capabilities, four dimensions of patient-level reasoning, three prerequisites for production-ready clinical agents, six governance requirements, and the FDA RWE design constraints.

This section is where the architecture responds. Patient Journey Intelligence was designed from the ground up to satisfy every one of those requirements, not as a collection of add-on features, but as a single, shared platform foundation that organizations build once and deploy across every secondary-use application.

Here's how the platform delivers on each of the requirements outlined above:



01 | Multimodal Data Integration into a Unified Living OMOP Foundation

Patient Journey Intelligence delivers all eight non-negotiable data capabilities as core platform functions:

- Multimodal ingestion handles structured EHR fields, clinical notes, scanned PDFs, imaging metadata, FHIR resources, DICOM files, lab results, and claims through a unified pipeline.

- Healthcare-specific SLMs extracts diagnoses, medications, findings, SDOH, and procedures from free text at 85–95% precision, approximately 30% more accurate than general-purpose LLMs on clinical tasks.
- All concepts are normalized to SNOMED CT, RxNorm, LOINC, and ICD-10-CM.
- Clinical reasoning resolves conflicts, deduplicates entities, and distinguishes confirmed diagnoses from ruled-out conditions.
- Longitudinal timelines organize every clinical event chronologically with precise temporal context.
- De-identification achieves 99%+ PHI removal under HIPAA and GDPR.
- Complete provenance tracks every derived fact from source document to OMOP output.
- The OMOP foundation updates continuously as new data arrives in order to ensure living datasets, not quarterly snapshots.

Multimodal Clinical Data Integration

Ingests structured **EHR** fields, clinical notes, scanned PDFs, imaging metadata, **FHIR** resources, and claims data through a unified pipeline.

NLP Medical Entity Extraction

Medical Language Models extract diagnoses, medications, and findings with 85–95% precision, understanding negation, uncertainty, and assertion status.

Terminology Standardization

All concepts normalized to **SNOMED CT, RxNorm, LOINC,** and **ICD-10-CM**, enabling accurate patient counts and cross-institutional research.

Clinical Reasoning

ML-based confidence scoring, entity deduplication, conflict resolution, and distinction between confirmed diagnoses vs. ruled-out conditions.

Longitudinal Timelines

Clinical events organized chronologically with temporal context for disease progression, treatment response, and time-to-event outcomes.

Privacy & De-Identification

99%+ accurate **PHI** removal in line with **HIPAA** and **GDPR** regulations, configurable for research and external data sharing.

Provenance & Auditability

Complete lineage from source document to **OMOP** representation with confidence scores and drill-down to supporting evidence.

Continuous Updates

Patient journeys update automatically as new data arrives, ensuring AI agents and analytics operate on current, complete representations.

02 | Patient Level Reasoning

The healthcare specific LLMs and SLMs included in the Patient Journey Intelligence Platform, together with the logic implemented by the tools and agents successfully deliver all four dimensions of patient-level reasoning as core extraction behavior, not optional post-processing:

- **Conflict resolution** identifies when the same condition is confirmed in one encounter and ruled out in another, and determines the authoritative assertion based on date, source type, and clinical context.
- **Temporal reasoning** tracks the diagnosis lifecycle from provisional to confirmed to resolved, preserving the progression in the longitudinal timeline rather than collapsing it to a single undated flag.
- **Assertion detection** understands negation, uncertainty, and family-history context, preventing ruled-out conditions, hypothetical findings, and relatives' diagnoses from being misattributed to the patient.
- **Deduplication** merges records of the same clinical entity across multiple documents into a single, attributed entry: one drug, one confidence score, one provenance chain.

03 | AI Ready Datasets

Patient Journey Intelligence provides the multimodal, reasoning-enriched foundation that agents need to operate on clinical truth rather than structured fragments. An agent querying the platform's patient timelines works on complete longitudinal representations, not the 13% of clinical concepts captured in structured fields.

All platform capabilities, patient data, cohort operations, terminology lookups, NLP extraction, document retrieval, and registry access are exposed via Model Context Protocol (MCP), the open standard for AI agent interoperability developed by Anthropic. Agents discover available tools dynamically, chain operations without hardcoded integrations, and compose workflows across platform functions and external services without bespoke API development.

Every agent action is captured in the same audit infrastructure that governs human data access, user identity, timestamp, data queried, logic applied, and output produced. Agents are constrained to authorized data by the same role-based access control enforced at the data level. Every recommendation an agent produces is explainable: traceable to source clinical facts, extraction method, confidence score, and the reasoning chain that connected them.

04 | Governance Requirements

De-identification and parallel datasets are handled by automatically maintaining synchronized identified and de-identified OMOP datasets from the same source. PHI removal uses HIPAA Safe Harbor methods extended with medical-specific patterns: date-shifting, consistent pseudonyms, tokenization, and structured/unstructured alignment, achieving 99%+ accuracy. Both datasets update continuously as new clinical data arrives, so research teams and operational teams never work from divergent versions of clinical truth.

Role-based access control is enforced at the data level, not just the application layer. Access policies scope by role, dataset, and use case, and AI agent permissions are governed by exactly the same controls as human users.

Audit logging captures every data access, NLP extraction, transformation step, and agent action in tamper-evident logs, recording user identity, timestamp, and clinical context. Logs are retained per institutional policy and retrievable in minutes, so when a regulator, or legal team asks who accessed what and when, the answer is already there, complete and verifiable, not reconstructed from memory.

Provenance and lineage flow end-to-end: from source document through NLP extraction, terminology normalization, OMOP mapping, and de-identification, every derived clinical fact carries its complete ancestry: document type, document date, extraction model, confidence score, and precise text location. Any output from a registry report, cohort query, or AI recommendation can be traced in a single step back to the original clinical note that supported it.

Secure deployment means the entire platform, including medical language models, runs within institutional infrastructure. No PHI flows to external services or shared cloud environments. Air-gapped deployments are fully supported. This boundary is enforced by architecture, not by configuration policy that a misconfiguration could undo.

Versioning and reproducibility ensure that every dataset state is time-stamped and preserved. Any result from any point in time can be exactly recreated using the data and model versions that produced it.

05 | FDA RWE Requirements

- **Relevance:** Healthcare-specific NLP extraction captures the clinical concepts that regulatory questions require: diagnoses, staging, biomarkers, adverse events, and SDOH, including the 87% of clinical concepts that exist only in free text and are invisible to structured-only approaches.
- **Reliability:** Medical language models achieve 85–95% extraction precision on clinical tasks, applied consistently across all document types and data sources, with published confidence scores for every extracted fact.
- **Completeness:** Multimodal ingestion and timeline completeness fills in the gaps for the 40% or less information available with structured-only approaches.
- **Traceability:** Every derived fact carries end-to-end provenance: source document, extraction model version, confidence score, OMOP mapping, and transformation log; with one-click drill-down to the original clinical text. The provenance chain is preserved automatically, not reconstructed manually.

What Patient Journey Intelligence Delivers - At a Glance

Data Completeness

Complete longitudinal timeline vs. 13% from structured fields alone. Multimodal ingestion across EHR, notes, PDFs, imaging, FHIR, and claims.

Clinical Accuracy

85–95% NLP extraction precision. Patient-level reasoning resolves conflicts, tracks diagnosis lifecycle, and correctly attributes assertions.

Regulatory Readiness

End-to-end provenance from source note to OMOP output. 99%+ de-identification. Versioned, reproducible results meeting 21 CFR Part 11 and FDA RWE guidance.

Agent-Ready Infrastructure

MCP-exposed APIs across all platform capabilities. Agents compose workflows without bespoke integrations. Full governance applies to every agent action.

Governance by Design

Role-based access at the data level. Tamper-evident audit logging. Private, secure deployment - architecture-enforced, not configuration-dependent.

Living Data Foundation

One shared OMOP dataset that all applications build on. Continuously updated. No redundant pipelines, no conflicting patient facts, no stale snapshots.

Build the foundation once. Innovate on accurate, trustworthy, reusable patient data - forever.

The result: instead of building incomplete pipelines that miss 40% of diagnoses, 81% of suicidality mentions, and 68% of cancer staging data, organizations build once on a foundation that captures what actually happened to patients, across all modalities, normalized to standard vocabularies, with full provenance and continuous updates.

What You Gain	Measurable Outcome
Speed to Insight	Timeline construction that took weeks of manual abstraction now completes in hours. Analyze 6× more patients in the same timeframe.
Completeness	>96% timeline completeness vs. 13% with structured-only approaches. One OMOP CDM format across all data.
Reuse at Scale	One standardized dataset supports cohorts, registries, analytics, and AI agents. Build the foundation once and innovate endlessly on top.
Embedded Governance	Full provenance, lineage, confidence scores, and audit trails for every derived clinical fact. Regulatory compliance built in.
Regulatory Readiness	Parallel identified and de-identified datasets, semantically synchronized. Research models move to production without pipeline rewrites.
Future-Proof AI	A durable foundation for advanced analytics, ML, and agentic systems. New AI applications added without re-engineering data preparation.

12. Shifting from Project-Specific Data Wrangling to a Scalable, Living Data Operating Model

The transformational shift Patient Journey Intelligence enables is not just cleaner data, it is a fundamentally different operating model for secondary use:

INSTEAD OF...	YOU GET...
80% effort on data wrangling per project	10% on data integration (once) and 90% on innovation
40% incompleteness accepted as the norm	>96% complete patient representations you can trust
Isolated pipelines that conflict and diverge	A single shared foundation consistent across all use cases
Quarterly snapshots that are stale on delivery	Living datasets updating automatically as new data arrives

Instead of spending 80% of effort on data wrangling for every new use case, teams spend 10% on data integration (once) and 90% on innovation (continuously).

Instead of accepting 40% incompleteness and hoping it doesn't matter, organizations build on 96%+ complete patient representations and trust their results.

Instead of isolated data pipelines that conflict and diverge, a single shared foundation ensures consistency across cohorts, registries, analytics, and AI applications.

13. The Mandate for Clinical Truth

Secondary use of clinical data has enormous potential, but only if the data accurately represents what actually happened to patients. Right now, for the vast majority of healthcare organizations, it does not. Read more on how Real-World Data Platforms handle the accuracy gap here.

The clinical data accuracy gap is not a hypothesis. It is documented in dozens of peer-reviewed studies, replicated across clinical data types, and measurable in the outputs of every analytics and AI initiative that relies exclusively on structured EHR data.

Patient Journey Intelligence provides the foundation to close this gap: complete multimodal data integration, healthcare-specific NLP achieving 85–95% extraction precision, terminology normalization, clinical reasoning, longitudinal patient timelines, 99%+ de-identification, full provenance, and continuous updates. Build the foundation once, and innovate on accurate, trustworthy, reusable patient data from that point forward.

Ready to close the gap?

Explore the Patient Journey Intelligence platform, review the full research evidence, and connect with the John Snow Labs team.

Get Started