

Scaling Specialist Hepatology Expertise for Population-Scale MASLD Care: An AI-Enabled OMOP CDM Knowledge Transport Layer for India

Ribhu Lahiri, Ganeev Singh, Kartik Mishra, Siddhant Janbandhu, Amit Goel, Vibhu Agarwal

Miimansa AI, Miimansa AI, Miimansa AI, Miimansa AI, SGPGIMS, Miimansa AI

Background

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects an estimated 250 million individuals in India, yet early detection is limited by the difficulty of delivering specialist hepatology expertise at scale.¹ MASLD management requires longitudinal interpretation of metabolic trajectories, laboratory measurements, fibrosis risk scores, and clinical context, while much of the population receives care in primary and frontline settings.^{2, 3}

Scaling MASLD care requires evidence-based hepatology workflows to operate beyond specialist centers. In many low- and middle-income country (LMIC) settings, however, clinically relevant evidence remains trapped in handwritten notes, scanned documents, laboratory reports, and fragmented records rather than standardized, computable form.

The OMOP Common Data Model (CDM) provides a community-maintained standard for representing clinical evidence.⁴ Although traditionally viewed as infrastructure for observational research, OMOP may also function as a knowledge transport layer, enabling validated disease models, risk scores, and evidence-based workflows to operate across heterogeneous care environments.

We present an AI-enabled pipeline that converts heterogeneous clinical documents into OMOP representations capable of supporting MASLD workflows outside specialist centers. Artificial intelligence functions not as the intervention itself, but as an enabling layer for standardized evidence generation at scale.

Methods

Data

Early development and validation used synthetic longitudinal patient records designed to resemble typical MASLD care pathways while avoiding use of identifiable patient data. Synthetic patients were generated from exemplar MASLD cases using LLMs to create clinically plausible discharge summaries, laboratory reports, outpatient encounters, and admission histories while preserving realistic relationships between biomarkers, medications, and disease progression.

These datasets supported early testing of OMOP-CDM transformations, longitudinal evidence reconstruction, and fibrosis risk score computation workflows while minimizing patient privacy risks. Clinical realism was assessed by a senior hepatologist who confirmed that disease trajectories and documentation patterns were representative of MASLD cases routinely encountered at the institution. The synthetic datasets generated and used during the current study are available from the corresponding author upon reasonable request.

Building a Standardized Evidence Infrastructure for MASLD Care

The primary challenge was converting scanned documents, handwritten notes, discharge summaries, and fragmented records into computable longitudinal representations for OMOP-native workflows. We developed an asynchronous NLP pipeline for semantic chunking, entity extraction, vocabulary standardization, patient-level sequencing, and ETL into OMOP CDM.

Three implementation learnings emerged. First, LLM-assisted chunking was useful for variable discharge summary formats. Second, vocabulary standardization was the main bottleneck: local drug brands, abbreviations, and spelling variants required localization before concept mapping. Third, patient-level sequencing before ETL was needed to resolve duplicate entities and reconstruct longitudinal evidence.

Row-Level DQD Traceability for Data Quality Remediation

Automated extraction from noisy source documents can introduce errors that propagate into downstream risk scores and workflows. Because aggregate data-quality reports do not identify which patients, encounters, or source documents require remediation, we extended the OHDSI Data Quality Dashboard (DQD)⁵ to enable row-level traceability from failed validation rules back to underlying clinical evidence (Figure 1).

This allowed reviewers to trace failures to patients, encounters, OMOP domains, and source documents while remaining compatible with future DQD releases. The main learning was that demographic linkage quality must precede remediation: when evidence is reconstructed from fragmented artifacts rather than a unified EHR, incomplete linkage prevents failed records from being corrected or reused for longitudinal reasoning.

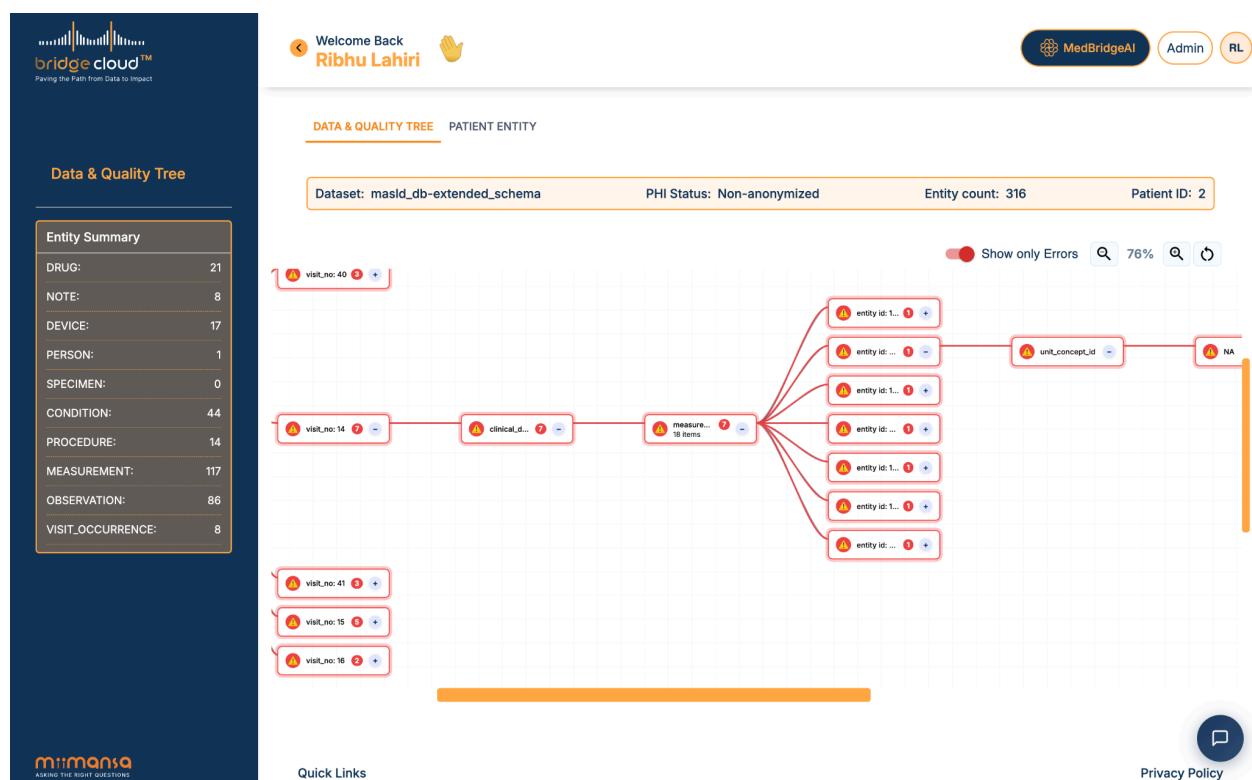


Figure 1. Row-level DQD traceability for OMOP data-quality remediation. Data-quality failures are displayed as traceable paths from patient to OMOP domain and field-level violation, allowing reviewers to locate the source of measurement errors such as missing or invalid `unit_concept_id` values.

Operationalizing Specialist Knowledge Through an OMOP-Native Clinical Intelligence Layer

OMOP functions as a knowledge transport layer only when standardized evidence supports reproducible workflows beyond CDM storage. We implemented a modular clinical intelligence layer that executes validated hepatology workflows directly against OMOP CDM representations, including FIB-4, AGILE 3+, AGILE 4, and SAFE fibrosis scores from standardized laboratory measurements.

Disease logic was separated from OMOP access infrastructure to maximize portability across institutions and allow new workflows without modifying the data access layer.

A physician-configurable dashboard surfaces fibrosis scores, longitudinal biomarker trajectories, and association views across metabolic, liver-function, and fibrosis-relevant measurements. Figures 2 and 3 show how OMOP-native evidence can be transformed into clinical representations that help non-specialist users identify patients whose trajectories, risk scores, or metabolic profiles warrant repeat evaluation or hepatology referral before advanced disease develops.



Figure 2. Longitudinal platelet trend with clinical-event context. The dashboard displays platelet count over time with reference limits, medication exposures, and fibrosis-related clinical events. A declining platelet trajectory approaching the lower reference threshold can help non-specialist users flag possible fibrosis risk and prioritize further evaluation.

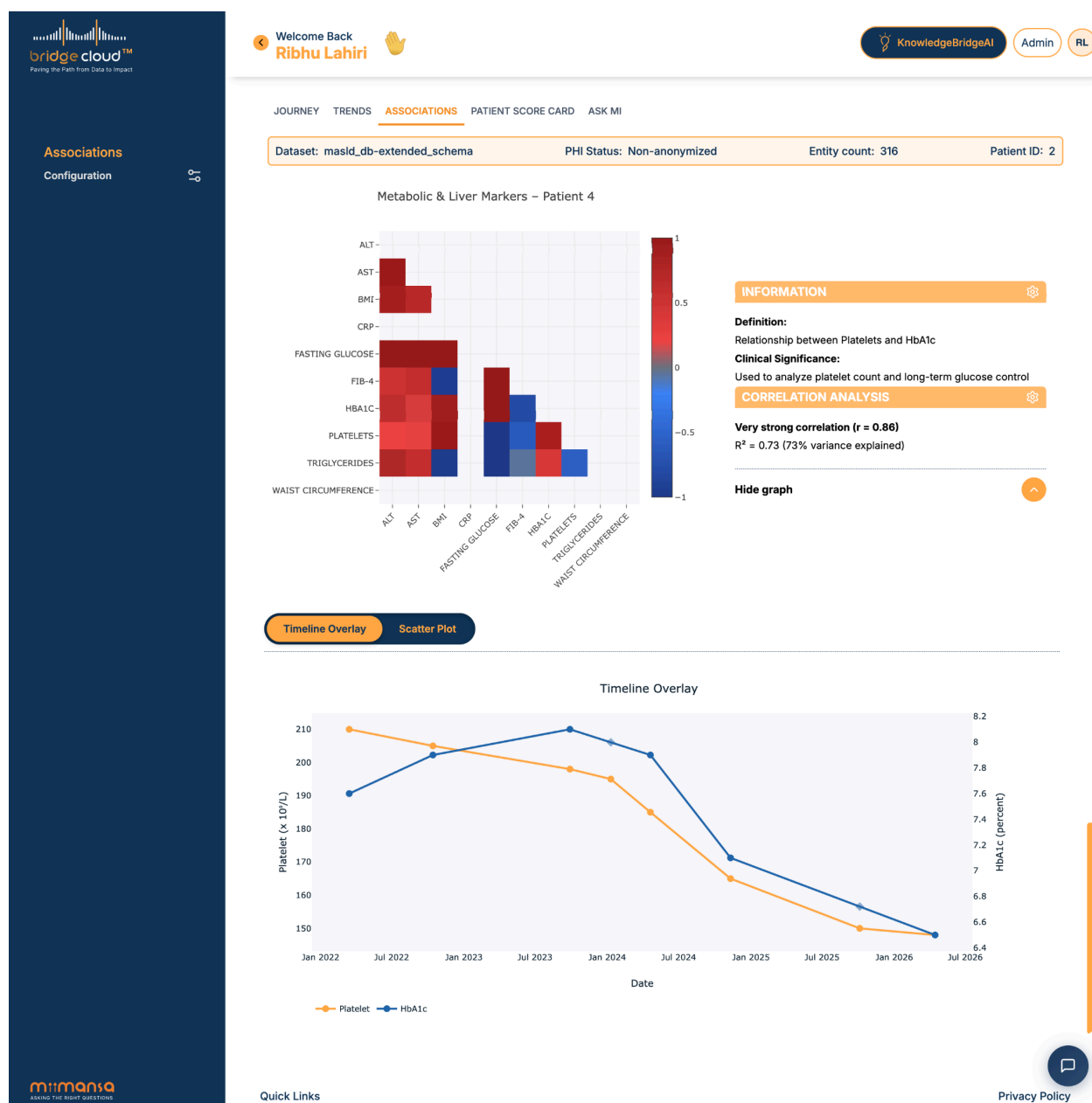


Figure 3. Metabolic and liver-marker association view with timeline overlay. The interface displays a correlation heatmap across MASLD-relevant variables and enables longitudinal comparison of selected markers, such as platelet count and HbA1c. This helps non-specialist users interpret fibrosis risk in the context of metabolic trajectory rather than relying on isolated lab values.

Results

Results are reported from an early synthetic evaluation dataset generated from exemplar MASLD care pathways and reviewed by a senior hepatologist at a partner institution for clinical plausibility. Table 1 summarizes feasibility metrics for OMOP transformation, entity extraction, vocabulary mapping, and row-level data-quality traceability.

Metric	Value
Documents processed	52
Document-source settings	1
Synthetic Patients	5
Synthetic Encounters	51
Extracted entities	1319
Standard concept mapping rate	85%
Entity extraction precision / recall	0.897 / 0.904
DQD checks run	2,043
DQD checks eligible for row-level traceability	1,641
Violations traceable to source row	80.03%

Table 1. Synthetic MASLD dataset evaluation summary. Metrics describe early feasibility testing of OMOP transformation, entity extraction, vocabulary mapping, and row-level data-quality traceability using hepatologist-reviewed synthetic longitudinal records.

Making Standardized Evidence Trustworthy and Actionable

Row-level DQD traceability attributed failed quality checks to patients, encounters, and source documents, enabling targeted remediation rather than aggregate failure reporting. Incomplete demographic linkage was an important source of traceability failure, showing that when evidence is reconstructed from fragmented records rather than unified EHR systems, linkage quality becomes a prerequisite for trustworthy longitudinal reasoning and downstream workflow execution.

Operationalizing Specialist Workflows Using OMOP-Native Evidence

We assessed computability and completeness for FIB-4, AGILE 3+, AGILE 4, and SAFE. These workflows executed when required laboratory measurements were available in the OMOP measurement layer, while missing laboratory values were the dominant cause of non-computability. This demonstrates the central implementation requirement for LMIC settings: portable specialist workflows depend not only on standardized representations, but also on reliable capture of the basic evidence needed to execute them. Figures 2 and 3 show how computable OMOP evidence can then be surfaced as longitudinal clinical views for non-specialist users.

Component	Contribution to Knowledge Transport	Transferable Learning
LLM-based NLP pipeline to OMOP CDM	Converts heterogeneous clinical evidence into standardized representations capable of supporting portable workflows	Context-aware chunking and localization were essential for heterogeneous clinical documentation
Row-level DQD traceability	Enables trustworthy standardized evidence through actionable remediation workflows	Linkage quality must precede meaningful quality assessment and remediation
OMOP-native clinical intelligence layer	Operationalizes specialist workflows using portable evidence representations	Disease logic should remain infrastructure independent to maximize portability
Federated OMOP deployment	Enables identical workflows across heterogeneous institutions without movement of raw patient data	Federated architectures improve deployability in regulatory and infrastructure-constrained settings

Table 2. Transferable lessons for OMOP-based knowledge transport. Components are summarized by their role in converting fragmented clinical evidence into portable workflows for heterogeneous, resource-constrained settings.

Conclusion

MASLD illustrates a broader LMIC challenge: clinically relevant evidence exists at population scale, while specialist expertise remains unevenly distributed relative to where patients receive care. Addressing this gap requires infrastructure that can transform routine clinical evidence into standardized representations capable of supporting validated workflows beyond specialist centers.

This MASLD use case demonstrates how OMOP CDM can function as a knowledge transport layer, not only as infrastructure for observational research. By converting fragmented clinical artifacts into standardized evidence and executing hepatology workflows against OMOP-native data, we present an implementation pattern for making specialist knowledge portable across heterogeneous care settings. In this framework, artificial intelligence is not the clinical intervention itself, but an enabling layer for standardized evidence generation in settings where data are fragmented, unstructured, and distant from specialist care.

References

1. Dieleman JL, et al. Global burden of disease and risk factors. *Lancet*. 2019.
2. Siddiqui MS, et al. Diagnosis of nonalcoholic fatty liver disease. *Am J Gastroenterol*. 2018;113(2):161–183.
3. Murugesan N, et al. Hepatology workforce in India: a survey. *J Clin Exp Hepatol*. 2021.
4. Hripcsak G, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574–578.
5. Kahn MG, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS*. 2016;4(1):1244.