TROVE HEALTH

WHITE PAPER

The Missing 70%

How Trove Recovers the Clinical Codes
That Interoperability Platforms Leave Behind

Venkat Timmaraju, PhD, MBA & Karthik Ravinutala
Trove Health Tech Inc.

March 2026

venkat@trovehealth.io | www.trovehealth.io

1. Executive Summary

Healthcare data interoperability remains one of the industry's most persistent challenges, costing the US healthcare system over $30 billion annually. At the heart of this problem lies a fundamental gap: the clinical data exchanged between health systems is overwhelmingly incomplete. Industry competitors such as Health Gorilla and Particle Health typically deliver only 30–40% of the clinical codes present in a patient's record, leaving critical diagnostic, procedural, and medication information buried in unstructured text and uncoded entries.

Trove Health has engineered a fundamentally different approach. Rather than competing solely on network access, Trove's differentiation lies in what happens after data retrieval: a multi-stage post-retrieval data processing pipeline that systematically discovers, extracts, and codes clinical information from every available data source—machine-readable CCD entries, human-readable CCD narratives, and unstructured clinical PDFs.

This white paper presents the results of a controlled pipeline run across 10 patients, demonstrating a 310% increase in clinical code yield over baseline coded data. These results validate Trove's core thesis: the real value in healthcare data interoperability is not merely in retrieving records, but in making every piece of clinical information within those records structured, coded, and actionable.

310%	16.5 min	962
Total Code Enrichment	Per-Patient Processing	Avg New Codes Per Patient

2. The Problem: Incomplete Clinical Data

When clinical data is exchanged between healthcare organizations via C-CDA documents and associated records, a significant proportion of the clinical narrative is never translated into standardized codes. This happens for several reasons: source EHR systems may not code all entries; human-readable narrative sections contain clinical detail that was never mapped to machine-readable fields; and a large volume of clinical documentation exists solely as scanned PDFs, entirely outside structured data flows.

The consequences are substantial. Incomplete coding leads to missed diagnoses in care gap analyses, inaccurate risk adjustment scores, denied claims, and degraded analytics. For organizations operating in value-based care models—ACOs, managed care plans, and risk-bearing provider groups—every uncaptured code represents lost revenue and compromised care quality.

2.1 Industry Baseline

Competitors in the healthcare data aggregation space focus primarily on network connectivity—connecting to health information exchanges such as CommonWell, CareQuality, and state HIEs

to retrieve patient records. While Trove also connects to these networks through its ConneXa platform (covering 99% of US clinical organizations), the critical difference is in data completeness.

Industry competitors typically deliver only the machine-readable, pre-coded portion of clinical records—approximately 30–40% of total available clinical information. Trove's pipeline processes all three data layers to achieve 98%+ data completeness.

2.2 The Three Layers of Untapped Clinical Data

Trove's pipeline identifies three distinct categories of data where clinical codes are absent, each requiring a different processing approach:

Machine-Readable (MR) Uncoded Data: Entries within the structured XML of C-CDA documents that contain clinical information but lack standardized codes (ICD, SNOMED, LOINC, RxNorm, etc.).

Human-Readable (HR) Uncoded Data: Narrative text sections within C-CDAs that describe clinical encounters, observations, and findings in natural language but were never linked to coded entries.

Clinical PDF Data: Scanned documents, discharge summaries, lab reports, and clinical notes delivered as PDF attachments—entirely unstructured and inaccessible to conventional aggregation platforms.

3. The Trove Data Enrichment Pipeline

Trove's pipeline is built on a principle of exhaustive extraction: every piece of clinical information in a patient's record should be identified, coded, and made available in structured FHIR R4 format. The pipeline operates across the three data layers described above, using a combination of proprietary parsing, database lookups, and TroveLLM—Trove's clinical large language model.

3.1 Machine-Readable (MR) Processing

C-CDA documents are first converted from XML into Trove's proprietary structured format. The pipeline then identifies entries in the machine-readable portion that lack standard clinical codes. These uncoded entries are processed through a two-stage approach:

Database Lookup: Entries are matched against Trove's comprehensive clinical terminology databases spanning 11 coding systems including ICD-10, SNOMED CT, LOINC, RxNorm, CPT, HCPCS, and NDC.

LLM-Powered Coding: Entries that cannot be resolved through direct database matching are passed to TroveLLM, which uses contextual understanding to identify the appropriate clinical codes.

Additionally, entries coded with non-standard or proprietary code systems are identified and mapped to recognized standards through database lookup.

3.2 Human-Readable (HR) Processing

Trove's proprietary data format enables a critical capability: linking human-readable narrative entries to their corresponding machine-readable counterparts. This linkage allows the pipeline to identify which HR entries are already represented in coded form (and therefore redundant) and which contain genuinely new clinical information.

After filtering out linked entries and cleaning rows with no useful clinical data, the remaining HR entries are sent to TroveLLM for entity extraction and code assignment. This approach ensures that only genuinely uncoded clinical narratives are processed, eliminating redundancy and maximizing the signal-to-noise ratio for the LLM.

3.3 Clinical PDF Processing

Clinical PDFs represent the largest untapped source of clinical information. Trove's PDF pipeline operates in multiple stages:

OCR Extraction: PDFs are processed through Trove's OCR pipeline to extract raw text from scanned documents, including handwritten clinical notes.

TroveLLM Segmentation: The extracted text is passed to the TroveLLM agent, which identifies relevant clinical text segments and classifies them into 17 HL7-standard sections (e.g., Problems, Medications, Procedures, Results, Allergies).

Code Discovery: Within each classified section, TroveLLM searches for and assigns appropriate clinical codes from the relevant coding systems.

This three-stage approach transforms completely unstructured documents into fully coded, section-classified clinical data that can be integrated directly into FHIR R4 resources.

4. Results: 10-Patient Pipeline Run

The following results are drawn from a controlled pipeline run processing clinical records for 10 patients. The baseline for comparison is the set of standard clinical codes present in the machine-readable portion of the original C-CDA documents.

4.1 Baseline

Initial standard codes in MR: 3,100 — This represents the starting point: the codes that would be available from a conventional data aggregation platform.

4.2 Machine-Readable Enrichment

MR Processing Category	DB Match	LLM Match	Total Added
Codes from uncoded MR entries	105	932	**1,037**
Codes from non-standard coded entries	437	0	**437**
Total MR codes added	**542**	**932**	**1,474**

Machine-readable processing alone added 1,474 new codes, representing a 47.5% increase over the baseline of 3,100 standard codes. Notably, the LLM resolved 90% of uncoded MR entries (932 of 1,037), demonstrating the importance of contextual understanding beyond simple database matching.

4.3 Human-Readable Enrichment

HR Processing Metric	Value
Total HR rows in records	45,952
HR rows linked to MR (redundant)	20,243 (44%)
Remaining unlinked HR rows	25,709
HR rows sent to LLM (after cleaning)	5,466
Codes generated from HR by LLM	**2,153**

The HR processing pipeline demonstrates the value of Trove's linking capability. Of the 45,952 human-readable rows, 44% were identified as duplicates of existing MR entries and excluded. After further cleaning of rows with no useful clinical data, 5,466 genuinely uncoded entries were processed by TroveLLM, yielding 2,153 new clinical codes—a 69.5% increase over baseline.

4.4 Clinical PDF Enrichment

PDF Processing Metric	Value
Average PDFs per patient	28.8
Average pages per patient	100.1
Total pages processed (10 patients)	~1,001
Codes added from PDFs	**5,992**

PDF processing produced the single largest contribution to code enrichment, adding 5,992 new codes—a 193% increase over baseline from this source alone. With an average of 28.8 PDFs and 100.1 pages per patient, this result underscores the enormous volume of clinical information that exists outside structured data exchange formats.

5. Cumulative Impact

The combined effect of processing all three data layers produces a transformative improvement in clinical code completeness:

Data Source	Codes Added	% Increase	Cumulative Total
Baseline (standard MR codes)	—	—	3,100
+ Machine-Readable enrichment	1,474	+47.5%	4,574
+ Human-Readable enrichment	2,153	+69.5%	6,727
+ Clinical PDF enrichment	5,992	+193%	12,719
TOTAL ENRICHMENT	**9,619**	**+310%**	**12,719**

From a baseline of 3,100 standard codes, Trove's pipeline added 9,619 new clinical codes—a 310% increase in total clinical code yield. The combined HR + MR processing alone (without PDFs) delivered a 117% improvement.

5.1 Source Contribution Analysis

The distribution of new codes across sources reveals an important insight: PDF documents contributed 62% of all newly discovered codes (5,992 of 9,619), followed by human-readable CCD narratives at 22% (2,153) and machine-readable uncoded entries at 15% (1,474). This distribution demonstrates that the majority of clinical information lost in conventional data exchange resides in unstructured documents that most aggregation platforms ignore entirely.

6. Performance, Economics, and Value

6.1 Processing Time

Pipeline Component	Time Per Patient
HR + MR coding pipeline	4.3 minutes
PDF processing pipeline	12.18 minutes
Total per patient (HR + MR + PDF)	**16.48 minutes**

At 16.48 minutes per patient for the complete pipeline, Trove can process approximately 87 patients per day on a single pipeline instance. The PDF pipeline accounts for 74% of processing time, reflecting the computational intensity of OCR extraction and LLM-based section classification across an average of 100 pages per patient. Pipeline instances scale horizontally, enabling processing of thousands of patients per day for population-level enrichment.

6.2 The Value Equation: Why 310% Enrichment Matters

The economic case for Trove's enrichment pipeline is best understood not through raw compute costs but through the downstream value of every clinical code recovered. In value-based care, risk adjustment, and quality measurement, missing codes translate directly into lost revenue, inaccurate risk scores, and failed quality measures.

Risk Adjustment (CMS-HCC): Each missed HCC code can represent $300–$3,000+ in annual per-member revenue. For a health plan managing 50,000 members, even a 2–3% improvement in code capture across the population can yield $5–15 million in additional risk-adjusted revenue annually.

HEDIS Quality Measures: HEDIS compliance depends on documented, coded evidence of care delivery. A missed medication code, a lab result buried in a PDF, or an uncoded diagnosis in a clinical narrative can be the difference between meeting and failing a quality measure. Since HEDIS looks back at the prior measurement year, Trove's pipeline is ideally suited to process the trailing 12 months of clinical data in a single enrichment pass—recovering codes that would otherwise count as gaps in care.

HEDIS Star Ratings Impact: For Medicare Advantage plans, each Star Rating increment can be worth $50–$100+ per member per month in quality bonus payments. When a plan manages hundreds of thousands of members, even fractional Star Rating improvements driven by better data completeness represent tens of millions in annual revenue.

Claim Denial Reduction: US hospitals lose an estimated $262 billion annually to claim denials, many of which stem from insufficient clinical documentation. Comprehensive coding from Trove's pipeline strengthens the clinical evidence supporting each claim, reducing denial rates and accelerating reimbursement.

6.3 ROI at Population Scale

To illustrate the value proposition concretely: consider a managed care organization with 100,000 members. If Trove's pipeline recovers even one additional HCC code for 10% of that population (10,000 members), at an average incremental value of $500 per recovered code, the annual revenue impact is $5 million. For HEDIS, closing even a small percentage of documentation-driven care gaps across the population can shift multiple quality measures from non-compliant to compliant, with cascading effects on Star Ratings and bonus payments.

> **The enrichment pipeline's value is measured not in compute dollars but in the millions of dollars of risk-adjusted revenue, quality bonuses, and avoided claim denials that flow from complete, coded clinical data. For HEDIS specifically, processing the prior measurement year's data in a single pass can recover the coded evidence needed to close care gaps at scale.**

6.4 Operational Efficiency

Beyond revenue impact, the pipeline eliminates the manual effort traditionally required to chase missing clinical documentation. Chart retrieval, manual abstraction, and retrospective coding

review are labor-intensive processes that health plans and ACOs spend millions on annually. Trove's automated pipeline replaces these manual workflows with a scalable, repeatable process that delivers results in minutes rather than weeks.

7. Technology Foundation

7.1 TroveLLM

The coding pipeline is powered by TroveLLM, Trove's proprietary clinical large language model. Built on a 33-billion parameter Mixture of Experts (MoE) architecture and fine-tuned on 5 million medical Q&A pairs, TroveLLM achieved 73.5% weighted accuracy across 7,202 medical questions from 10 standardized benchmarks—outperforming Google's MedGemma by 14.2 percentage points.

TroveLLM's architecture is uniquely suited to clinical coding tasks: its expert subnetworks specialize in distinct medical domains (pharmacology, diagnostics, procedures), and its training on data preprocessed by Trove's 35-model Trident AI parsing system gives it an inherent advantage in understanding the structure and semantics of clinical documents.

7.2 Trident AI

Trident AI is Trove's 35-language-model ensemble parsing system, achieving 98% accuracy in processing C-CDA documents and clinical PDFs. It processes over 11 million files monthly at an average of 1.07 seconds per document. Trident AI's output—perfectly structured, high-quality clinical data—serves as both the input to the enrichment pipeline described in this paper and the training data foundation for TroveLLM, creating a compounding data quality advantage.

7.3 ConneXa

ConneXa is Trove's FHIR-native interoperability platform, providing access to 99% of US clinical organizations through integrations with CommonWell Health Alliance, CareQuality Framework, 45+ state health information exchanges, and direct EHR APIs (Epic, Cerner, Allscripts, athenahealth). ConneXa serves as the data acquisition layer that feeds records into the Trident AI and enrichment pipeline.

8. Implications for Healthcare Stakeholders

8.1 Digital Health Companies

For digital health platforms that rely on patient data to drive clinical workflows, the difference between 30% and 98%+ data completeness is the difference between guessing and knowing. Trove's enrichment pipeline ensures that care gap analyses, medication reconciliation, and clinical decision support tools operate on comprehensive coded data rather than partial records.

8.2 Payers and Risk-Bearing Organizations

Accurate and complete clinical coding is the foundation of risk adjustment under CMS-HCC, HEDIS quality measurement, and value-based payment models. The 310% code enrichment demonstrated in this analysis directly translates to more accurate risk scores, higher quality ratings, and reduced claim denials. For a typical health plan, the revenue impact of even a modest improvement in coding completeness can reach millions of dollars annually.

8.3 Accountable Care Organizations

ACOs operating under MSSP or other shared savings models depend on comprehensive clinical data to identify care gaps, stratify risk, and demonstrate quality improvement. Trove's pipeline ensures that the clinical intelligence feeding these programs reflects the full scope of patient health—not just the fraction that was coded at the source.

9. Conclusion

The results presented in this white paper demonstrate a fundamental truth about healthcare data interoperability: retrieval is only half the problem. The real challenge—and the real value—lies in transforming the incomplete, inconsistently coded records that flow through health information exchanges into comprehensive, standardized clinical intelligence.

Trove's multi-source enrichment pipeline achieves this transformation at scale, delivering a 310% increase in clinical code yield with processing times measured in minutes per patient. By processing all three layers of clinical data—machine-readable, human-readable, and unstructured PDFs—the pipeline captures clinical information that conventional aggregation platforms systematically miss. For HEDIS and risk adjustment use cases, the ability to enrich a full measurement year of data in a single automated pass replaces costly manual chart retrieval and abstraction workflows.

For healthcare organizations seeking to build on a foundation of complete, accurate clinical data, Trove Health offers not just a data platform but a data quality advantage that compounds over time—better data driving better AI, driving better outcomes.

Learn More

Venkat Timmaraju, PhD, MBA & Karthik Ravinutala
venkat@trovehealth.io | www.trovehealth.io