TROVE HEALTH

WHITE PAPER

One Patient, One Record

How Trove's Probabilistic EMPI Engine Achieves 99.55% Match Accuracy Across 21 Million Clinical Documents

Vinay Miryala & Venkat Timmaraju, PhD, MBA
Trove Health Tech Inc.

March 2026

venkat@trovehealth.io | www.trovehealth.io

1. Executive Summary

Patient identity is the foundation of every clinical data operation in healthcare. When a patient's records are fragmented across multiple identities—or worse, when two different patients are merged into one—every downstream process is compromised: risk adjustment scores are inaccurate, care gap analyses are unreliable, and clinical decision support operates on an incomplete picture.

Trove Health has built a next-generation Enterprise Master Patient Index (EMPI) engine that combines probabilistic scoring with intelligent rule-based gating to solve this problem at scale. Unlike conventional deterministic matching systems that require exact field agreement, Trove's EMPI uses weighted multi-field similarity scoring, cultural name intelligence, and false-positive gate rules to handle the messy reality of healthcare demographics—where names are misspelled, dates of birth are transposed, addresses are outdated, and the same person appears under a nickname, a maiden name, or a cultural variant.

This white paper presents the design principles and production results of the EMPI Scoring Engine v4.0. The system has been validated across over 21 million clinical documents spanning 136,000+ patients in production, achieving a 76.69% auto-match rate, a combined 99.55% resolution rate (auto-match plus probable-match), and a no-match rate of just 0.45%.

21M+	99.55%	0.45%
Documents Processed	Resolution Rate	No-Match Rate

2. The Problem: Fragmented Patient Identity

Healthcare data is generated across dozens of systems—EHRs, laboratory information systems, radiology platforms, pharmacy management systems, and health information exchanges—each maintaining its own patient registration process. The same patient may be registered as "Robert Smith" in one system, "Bob Smith" in another, and "Robert J. Smyth" in a third, with slight variations in date of birth, address, and other demographics.

The scale of this problem is staggering. Studies consistently estimate that 8–12% of patient records in the average health system are duplicates, and cross-organizational matching rates are significantly worse. When healthcare organizations exchange data through interoperability networks—CommonWell, CareQuality, state HIEs—the patient identity problem is amplified: records from external systems must be matched against the receiving organization's master patient index with no guarantee of shared identifiers.

2.1 The Cost of Getting It Wrong

Patient identity errors create two categories of failure, each with severe consequences:

False negatives (missed matches): When the system fails to recognize that two records belong to the same patient, the patient's clinical history is fragmented. Clinicians see an incomplete picture, risk adjustment misses documented conditions, and care gap analyses undercount services delivered. For organizations processing data through Trove's enrichment pipeline—which can discover 310% more clinical codes from a patient's records—a missed identity match means those enriched codes are attributed to the wrong identity or lost entirely.

False positives (incorrect merges): When two different patients are incorrectly linked, clinical data is co-mingled. A medication allergy from Patient A appears on Patient B's record. A diagnosis code from one family member is attributed to another. In risk adjustment, false positives can trigger compliance violations and audit exposure. In clinical care, they are a patient safety risk.

2.2 Why Deterministic Matching Fails

Traditional EMPI systems rely on deterministic rules: if first name, last name, date of birth, and SSN all match exactly, the records are linked. This approach fails in practice because healthcare demographic data is inherently noisy. Names are abbreviated ("Md" for "Mohammed"), misspelled ("Smyth" for "Smith"), or changed (maiden names). Dates of birth are transposed (03/15 entered as 15/03). Addresses are outdated. Social Security Numbers are frequently missing, incorrect, or intentionally withheld.

A system that demands exact agreement on these fields will miss a large proportion of legitimate matches—precisely the records that need to be linked for complete clinical intelligence.

3. Trove's EMPI: Architecture and Design

Trove's EMPI engine takes a fundamentally different approach: probabilistic scoring combined with intelligent gating. Every candidate pair receives a continuous similarity score between 0.0 and 1.0, computed from weighted field-level comparisons. Rule-based gates then prevent specific categories of false positives from being promoted to matches, even when the aggregate score is high.

3.1 Scoring Fields and Weights

The engine evaluates five demographic fields, each independently scored on a 0.0–1.0 scale and combined using normalized weights:

Field	Weight	Role in Matching
First Name	**~29.5%**	Highest weight; variant expansion, nicknames, cultural aliases
Last Name	**~23.6%**	Compound surname handling, phonetic matching
Date of Birth	**~23.6%**	9-rule priority cascade for typos, transpositions

Address	**~11.8%**	USPS normalization, fuzzy rescue, PO Box handling
ZIP Code	**~11.8%**	Two-tier gate: conditional on address quality

Gender is scored separately but excluded from the weighted probability formula. It serves as a rescue mechanism: when all other identity signals are strong but the aggregate score falls below threshold due to a missing address, gender agreement can rescue the match to the probable-match tier.

3.2 Three-Tier Match Classification

The final weighted score drives a three-tier classification:

Classification	Score Range	Action
Auto-Match	≥ 0.96	Records linked automatically — no manual review required
Probable-Match	0.80 – 0.9599	Flagged for human review — moderate confidence
No-Match	< 0.80	Records not linked — insufficient evidence

The 0.96 auto-match threshold is deliberately conservative. In healthcare, the cost of a false positive (merging two different patients) is far greater than the cost of a false negative (requiring manual review). By setting the bar high for automatic linking and providing a probable-match tier for human adjudication, the system optimizes for safety while still resolving the vast majority of records without manual intervention.

4. Built-In Clinical and Cultural Intelligence

What distinguishes Trove's EMPI from conventional matching systems is the depth of domain-specific intelligence built into every scoring function. The engine doesn't just compare strings—it understands the patterns and variations that healthcare demographic data actually exhibits.

4.1 Name Intelligence

Nickname Resolution: The engine uses bidirectional nickname matching to resolve common name variants. "Bob" matches "Robert," "Peggy" matches "Margaret," and "Bill" matches "William." Nickname matches score 0.95 (rather than 1.0) to distinguish them from verified identity matches while still contributing strongly to the overall probability.

Cultural Name Variants: The system includes dedicated alias tables for cultural naming patterns that string-similarity algorithms cannot resolve. "Md"—a standard South Asian short form—is recognized as equivalent to "Mohammed," "Mohammad," "Muhammad," and other variants. Without this table, Jaro-Winkler similarity between "Md" and "Mohammed" would score approximately 0.52—a near-certain miss.

Compound Surname Handling: The engine generates no-space variants for compound surnames, ensuring that “Diaz Cosme” (spaced) matches “Diazcosme” (concatenated)—a common artifact of EHR data entry.

Phonetic Matching: Double Metaphone encoding catches spelling variations that preserve pronunciation: “Smith” and “Smyth” share the same phonetic key and score 1.0.

4.2 Date of Birth Intelligence

Date of birth is one of the strongest identity signals in healthcare, but it is also prone to specific categories of data entry error. Rather than treating DOB as binary (match or no-match), Trove’s engine applies a 9-rule priority cascade that recognizes common error patterns and assigns graduated scores:

Rule	Score	What It Catches
Exact match	1.00	Identical dates
Day/month swap	0.85	MM/DD vs DD/MM entry (03/15 ↔ 15/03)
Within 2 days	0.80	Off-by-one or off-by-two day entry errors
Year differs by 1	0.60	Year boundary errors (1984 vs 1985)
Year digit transposition	0.50	Keyboard transposition (1987 vs 1978)
Day digit transposition	0.45	Day reversal (13 vs 31)
Single-digit typos	0.20	Keystroke errors in year or day

This graduated approach is critical. A day/month swap—common when data is entered in international date formats—is far more likely to represent the same patient than a completely different date. The cascade captures this nuance quantitatively.

4.3 Address Intelligence

Address comparison incorporates USPS abbreviation equivalence (“Street” ↔ “St,” “Boulevard” ↔ “Blvd”), directional normalization (“North” ↔ “N”), unit keyword unification (“Apartment,” “Suite,” “Unit” → “Apt”), and ordinal mapping (“First” ↔ “1st”). A fuzzy token rescue mechanism matches misspelled street names (Jaro-Winkler ≥ 0.92 on tokens of 4+ characters), and a street-number anchor bonus rewards matches that share both a house number and a meaningful street name fragment.

The engine also handles PO Box addresses via a dedicated fast-path that compares only the box number, avoiding the false similarity inflation that occurs when a generic “POBOX” token dominates string-level comparisons.

4.4 Gender-Based Fallback Rescue

The system includes a gender-aware fallback mechanism that recognizes a pattern specific to healthcare: a patient whose name and date of birth match perfectly, but whose address is missing or outdated. For male patients, the fallback requires exact agreement on first name, last name,

date of birth, and gender. For female patients, the last name requirement is relaxed to accommodate maiden name changes—a design decision grounded in the reality that women frequently appear under different surnames across healthcare systems at different life stages.

5. False-Positive Prevention: Gate Rules

High aggregate scores do not always indicate a true match. Family members sharing a household often have the same last name, address, and even date of birth (twins), differing only in first name. Without protection, these records would score highly and be incorrectly merged.

Trove's gate rule system prevents seven specific categories of false positives:

Gate Rule	What It Prevents
Same Household	Blocks merging family members who share DOB, last name, and address but have different first names
Diff First Name + Address	Blocks when first name differs and address provides no corroboration
Diff First Name	Blocks when first name differs even with similar address but inconclusive ZIP
Diff Last Name + Address	Blocks when last name and address both differ despite matching first name and DOB
Diff Last Name	Blocks when last name clearly differs, even at the same address
Diff First + Last Name	Blocks when both names differ despite strong DOB match
Diff Address + Gender	Blocks when strong name+DOB is present but all location and gender signals fail

When any gate rule fires, the match probability is capped at 0.799—just below the probable-match threshold—and the gate label is recorded in the output for auditability. Gate rules are evaluated in priority order, with the first firing gate taking precedence. This layered approach ensures that the system is transparent about why a potential match was blocked, supporting the manual review workflows that healthcare organizations depend on.

Gate rules are the safety net that makes high auto-match rates possible. By proactively identifying and blocking the most common false-positive patterns—household members, twins, maiden name collisions—the system can set an aggressive auto-match threshold (0.96) while maintaining patient safety.

6. Production Results

The EMPI Scoring Engine v4.0 has been validated in production across Trove's full patient document corpus. The following results reflect actual matching outcomes, not synthetic benchmarks.

6.1 Overall Match Distribution

Category	Documents	% of Total
Auto-Match (≥ 0.96)	16,211,462	**76.69%**
Probable-Match (0.80 – 0.9599)	4,736,311	**22.86%**
No-Match (< 0.80)	93,903	**0.45%**
TOTAL	**21,041,676**	**100.00%**

76.69%	22.86%	0.45%
Auto-Matched (No Review Needed)	Probable-Match (Review Queue)	No-Match (Unresolved)

6.2 Confidence Distribution

The distribution of scores across probability bands reveals the engine's confidence profile. Over half of all documents (51.68%) score above 0.98, indicating near-perfect demographic agreement. The auto-match tier (above 0.96) captures 76.69% of all documents across 107,980 unique patients.

Probability Band	Documents	% of Total	Classification
> 0.98 – 1.00	10,873,533	51.68%	**Auto-Match**
> 0.96 – 0.98	5,263,041	25.01%	**Auto-Match**
> 0.90 – 0.96	2,474,058	11.76%	**Probable-Match**
> 0.80 – 0.90	2,262,253	10.75%	**Probable-Match**
< 0.80	93,903	0.45%	**No-Match**

The extremely low no-match rate (0.45%) demonstrates the engine's effectiveness at resolving patient identity even when demographic data is imperfect. The majority of records in the probable-match tier (0.80–0.96) represent cases with one or two partial field disagreements—a nickname instead of a legal name, a transposed date digit, or an outdated address—that are readily adjudicated by human reviewers.

99.55% of all clinical documents are resolved to a patient identity—either automatically (76.69%) or with moderate confidence for review (22.86%). Only 0.45% of documents remain truly unresolved.

7. Deployment and Performance

The EMPI engine is deployed as an AWS Lambda function, providing serverless scalability that automatically adjusts to document processing volume. Each scoring invocation is stateless and self-contained, enabling horizontal scaling without infrastructure management.

The engine supports two operational modes. In block mode (the default for standard deduplication), the engine scores each document only against its own patient's existing records, grouped by a (document_id, patient_id) tuple to prevent cross-patient data pollution. In full cross-join mode (used for global linkage scenarios), every original record is compared against every document for comprehensive identity resolution across the entire patient index.

Each Lambda invocation returns a complete scoring breakdown—field-level scores, match classification, gate rule triggers, gender fallback indicators, and nickname resolution details—providing full transparency into every match decision for audit and compliance purposes.

8. The EMPI in Trove's Integrated Platform

Patient identity resolution is not an isolated function at Trove—it is the foundational layer upon which the entire data intelligence platform depends. The EMPI sits at the front of Trove's processing pipeline, ensuring that every clinical document is attributed to the correct patient before any downstream analysis begins.

8.1 ConneXa: Data Acquisition

ConneXa, Trove's FHIR-native interoperability platform, connects to 99% of US clinical organizations through CommonWell Health Alliance, CareQuality Framework, 45+ state health information exchanges, and direct EHR APIs. When ConneXa retrieves clinical records from these networks, the EMPI engine is the first processing step: every incoming document is scored against the master patient index to establish identity before any clinical processing occurs.

This is where the EMPI's quality directly impacts everything downstream. A misidentified document means that clinical codes, care gap evidence, and risk adjustment data are attributed to the wrong patient—or lost entirely. With a 99.55% resolution rate across 21 million documents, the EMPI ensures that ConneXa's broad network reach translates into accurately attributed clinical data.

8.2 Trident AI: Parsing and Structuring

Trident AI, Trove's 35-language-model ensemble parsing system, processes over 11 million files monthly at 98% accuracy, converting C-CDA documents and clinical PDFs into structured, coded data. Trident AI depends on correct patient attribution: when it processes a clinical document, it must know which patient the structured output belongs to. The EMPI provides this identity resolution at the point of ingestion, before Trident AI begins its parsing work.

8.3 TroveLLM: Clinical AI and Code Enrichment

TroveLLM, Trove's clinical large language model, powers the code enrichment pipeline described in Trove's "The Missing 70%" white paper—achieving a 310% increase in clinical code yield by processing machine-readable, human-readable, and unstructured PDF data. Every one of those 9,619 newly discovered codes (per 10-patient cohort) must be attributed to the correct patient identity for the enrichment to have value. A 310% code enrichment applied to a misidentified patient is worse than no enrichment at all.

The EMPI is not a standalone utility—it is the identity backbone of the entire Trove platform. ConneXa retrieves records, the EMPI resolves identity, Trident AI structures the data, and TroveLLM enriches it. Each layer depends on the accuracy of the one before it. At 99.55% resolution across 21 million documents, the EMPI ensures that Trove's clinical intelligence is built on a verified patient identity foundation.

8.4 Impact on Risk Adjustment and HEDIS

For organizations using Trove's platform for risk adjustment and HEDIS quality measurement, the EMPI's accuracy has direct financial consequences. With CMS now calculating 100% of Medicare Advantage risk scores under the V28 HCC model, every HCC code must be attributed to the correct patient to be submitted for risk adjustment. Similarly, HEDIS measures require documented, coded evidence of care delivery linked to a specific member identity. A missed match means missed revenue; a false merge means compliance risk.

Trove's EMPI eliminates identity uncertainty from these high-stakes processes, ensuring that the enriched clinical data flowing from the platform is both comprehensive and correctly attributed.

9. Conclusion

Patient identity resolution is the silent prerequisite for every clinical data operation in healthcare. Without it, interoperability is just data movement. With it, clinical records become complete, attributed, actionable intelligence.

Trove's EMPI Scoring Engine v4.0 represents a generational advance over deterministic matching: probabilistic scoring that handles the messy reality of healthcare demographics, cultural intelligence that resolves naming patterns invisible to conventional algorithms, and gate rules that prevent the false positives that erode trust in automated systems. Validated across 21 million clinical documents with a 99.55% resolution rate, it provides the identity foundation that makes Trove's entire data platform possible.

For healthcare organizations seeking to build clinical intelligence on a foundation of accurate, verified patient identity, the EMPI is where it all begins.

Learn More

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