ClinFuse: Patient Entity Resolution Powered by MedGemma Clinical Reasoning

Alexander Rider

1. Problem Statement

A cardiac arrest patient's record was confused with another patient's—one who carried a do-not-resuscitate order. The care team withheld life-saving treatment [1]. When healthcare systems can't reliably match records to the right patient, people die. Patients with duplicate charts—the most common form of this identity failure—are five times more likely to die during hospitalization [2]. Across 7,613 wrong-patient events at 181 healthcare organizations, 9% resulted in patient harm [1]. For the emergency department registrar matching an unconscious patient to their records, or the IT staff reconciling patient feeds from dozens of hospitals, these are not edge cases—they are daily realities.

Duplicate records aren't just dangerous—they're pervasive. An estimated 8–12% of hospital records are duplicates [3], costing the U.S. healthcare system $6.7 billion annually [4]. Match accuracy averages 80–90% within institutions but falls to 50% in cross-organizational exchange [5]—meaning roughly half of cross-organizational duplicates go unresolved. The MATCH IT Act (H.R. 2002, 2025) sets a 99.9% matching target [7], and TEFCA v2.1 makes cross-organizational matching critical national infrastructure [8]. Closing this gap would resolve 99.8% of currently-missed cross-organizational duplicates (from 50% unresolved to 0.1%). Even conservatively attributing half of the $6.7B burden to cross-organizational failures, this represents over $3 billion in recoverable annual waste—alongside the patient harm that fragmented records cause.

The barrier isn't better algorithms—it's better information. Any method relying exclusively on demographic fields faces an accuracy ceiling imposed by field quality—over half of duplicate pairs contain misspelled names or mismatched identifiers [3]. The missing signal is *clinical context*: two records sharing the same disease trajectory, medication regimen, and vital sign patterns are almost certainly the same patient, regardless of demographic discrepancies—but interpreting clinical histories as identity evidence requires medical reasoning that traditional matching systems cannot perform.

2. Overall Solution

To demonstrate MedGemma's potential for patient entity resolution, we embed it in ClinFuse, a three-tier pipeline that augments fast probabilistic matching with MedGemma's clinical reasoning, invoked selectively for ambiguous cases. Figure 1 illustrates the architecture.

Tier 1: Probabilistic Screening. All records enter Splink v4 [13] for Fellegi–Sunter [9] probabilistic linkage using expectation-maximization on demographic fields with Jaro–Winkler string similarity and term-frequency adjustment. Pairs with high match probability are auto-matched; pairs with very low probability are auto-rejected.

Tier 2: Clinical LLM Classification. Ambiguous "gray zone" pairs—those where demographics alone are inconclusive—receive structured clinical summaries and are classified by a fine-tuned MedGemma 4B model, whose medical pretraining enables it to interpret clinical histories as identity evidence. Each patient's clinical history is summarized into a compact format: conditions grouped by onset year, medications with date ranges, sorted allergies, latest values per vital sign, and procedures with years. The model sees two parallel summaries and classifies whether they describe the same patient (Figure 2).

Tier 3: Score Fusion. The probabilistic match score and the LLM's classification logit are combined in log-odds space using interpretable linear weights. A demographic veto floor prevents the LLM from overriding cases where demographics strongly disagree, guarding against false merges—the highest-risk failure mode in clinical settings. This interpretable design ensures each component's contribution is transparent and auditable.

Golden Records. Matched pairs form a graph whose connected components yield patient clusters. Field-level conflict resolution applies majority voting with domain heuristics (e.g., preferring longer address forms), producing a deduplicated Master Patient Index (MPI).

MedGemma as Clinical Reasoner. Prior work has applied LLMs to adjacent healthcare tasks—MedLink uses diagnosis-code embeddings for de-identified record linkage [10], and PRISM fine-tunes an LLM for clinical trial matching [11]—but neither addresses patient-to-patient identity resolution using clinical narratives. In general-domain entity resolution, large language models such as GPT-4 achieve strong zero-shot matching that rivals fine-tuned pre-trained language models [20], and cost-efficient frameworks selectively route only uncertain pairs to an LLM [21, 22]—an architecture ClinFuse shares. However, these approaches rely on cloud-hosted models with 70B–175B+ parameters, incurring per-query costs, introducing latency, and—critically for healthcare—requiring patient data to leave the facility, making them infeasible under HIPAA and GDPR. Beyond privacy, the scale of the matching problem itself rules out cloud approaches: matching 1 million records requires evaluating nearly 500 billion pairwise comparisons—a volume that makes per-query API calls infeasible on cost and latency alone, and demands a system that screens cheaply at scale and reserves clinical reasoning for only the cases that need it. Fine-tuned models of comparable size (e.g., Mistral-7B-Instruct, Qwen-14B) lack medical domain knowledge and underperform on clinical text [11]. A recent healthcare-specific approach fine-tunes PubMedBERT with contrastive learning for patient linkage at scale [23], and is deployable on-premises, but operates exclusively on demographic fields without leveraging clinical context.

MedGemma is the first medical foundation model that combines clinical reasoning with edge-deployable size, making it uniquely suited for healthcare entity resolution where privacy constraints rule out cloud models. Pretrained for clinical question answering and medical image interpretation, patient entity resolution is entirely outside its pretraining distribution—yet its

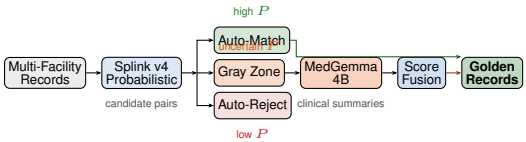

Figure 1: ClinFuse three-tier architecture. High-confidence pairs are auto-matched (green) or auto-rejected (red). Only ambiguous gray-zone pairs (amber) are routed through the clinical LLM, keeping compute costs low.

medical language understanding enables it to recognize that "Essential Hypertension" and "Hypertension" are the same condition, or that overlapping medication regimens constitute strong identity evidence. At 4B parameters with 4-bit quantization, it runs on a single consumer GPU (~$1,500), requires no cloud connectivity, and its medical pretraining provides the clinical reasoning necessary to interpret disease trajectories, medication overlaps, and vital sign concordance as identity evidence. QLoRA fine-tuning successfully adapts it to pairwise classification, achieving 0.93 F1 on held-out evaluation data—demonstrating that medical pretraining transfers effectively to this novel task.

3. Technical Details

3.1 Data Generation and Augmentation

Synthea [14] generates realistic synthetic patients with clinically coherent histories spanning conditions, medications, allergies, observations, and procedures. An augmentation pipeline distributes each patient's records across multiple simulated facilities and injects demographic errors: name variations (nickname substitution, typos, maiden name usage), address errors (abbreviation, format variation), date perturbations, identifier errors (SSN transposition, digit substitution), and formatting noise—mirroring documented real-world error patterns [3]. Separate datasets are generated for training (10,000 patients) and evaluation (500 patients, 1,275 records across 5 facilities).

3.2 Adapting MedGemma for Entity Resolution

Text-only adaptation. MedGemma 4B is a multimodal model with both vision and text encoders. Since our task is text-only, we strip the vision tower before fine-tuning, reducing from 4.2B to 3.88B parameters while preserving the medical language understanding from pretraining.

QLoRA fine-tuning with classification head. We apply 4-bit QLoRA [15] (Low-Rank Adaptation [16]) targeting attention and MLP projection layers, and add a binary classification head. Fine-tuned classification heads significantly outperform zero-shot prompting for binary tasks [17] while being orders of magnitude faster at inference [18]. Training uses balanced match/non-match pairs generated from augmented Synthea records. The fine-tuned adapter and merged model are published on HuggingFace Hub.

3.3 Results

Table 1 compares Splink-only matching against the full ClinFuse pipeline on 1,275 synthetic records (500 patients across 5 facilities).

At the conservative 0.95 threshold, Splink alone achieves perfect precision but misses true matches in the gray zone. MedGemma's clinical understanding—recognizing matching disease trajectories and overlapping medication regimens—recovers all missed matches, lifting recall from 0.991 to 1.000. The trade-off is a small precision reduction (6 false merges out of 1,187 predicted matches)—acceptable given the far higher clinical cost of missed matches. The demographic veto floor prevents the LLM from overriding strong demographic disagreement, keeping false merges rare. The MedGemma classifier achieves 0.93 F1 in standalone evaluation; MedGemma provides clinical understanding that no probabilistic linker can offer, complementing Splink's high-precision demographic matching.

Table 1: Entity resolution performance. Splink-only uses the 0.95 auto-match threshold. ClinFuse adds MedGemma gray-zone classification and score fusion.

Metric	Splink Only	ClinFuse
Precision	1.000	0.995
Recall	0.991	**1.000**
F1	0.995	**0.997**
Operational statistics (ClinFuse):		
Gray-zone pairs classified by LLM		242
LLM-recovered matches		4

Record A (Facility 1)

```
CONDITIONS:
  2018:  Hypertension *; Type 2 Diabetes *
  2020:  Acute Bronchitis
MEDICATIONS:
- Metformin 500mg (2018-ongoing)
- Lisinopril 10mg (2018-ongoing)
OBSERVATIONS:
- A1c:  7.2% (2024-06), 7.0% (2023-12)
```

Record B (Facility 3)

```
CONDITIONS:
  2018:  Essential Hypertension *; Diabetes *
  2020:  Acute Bronchitis
MEDICATIONS:
- Metformin Hydrochloride 500mg (2018-ongoing)
- Lisinopril 10mg (2019-ongoing)
OBSERVATIONS:
- A1c:  7.2% (2024-06), 7.0% (2023-12)
```

Figure 2: Structured clinical summary pair as input to the MedGemma classifier. Despite demographic discrepancies (name typo + address abbreviation), the parallel clinical trajectories—identical chronic conditions, overlapping medications, and matching vitals—enable the model to correctly identify a match.

3.4 Deployment and Feasibility

Hardware. MedGemma 4B runs quantized (4-bit NF4) on a single consumer GPU (e.g., RTX 3090, ~$1,500). No cloud API dependency.

Privacy by design. All inference is local—no patient data leaves the facility. This satisfies HIPAA and GDPR requirements and supports air-gapped deployment.

Cost at scale. Splink handles ~95% of pairs in under 1 second per pair. Only ~5% hit the LLM. At hospital scale (100K records), estimated total GPU time is under 1 hour. The pipeline supports both batch processing for MPI deduplication and near-real-time matching for point-of-care patient registration.

Integration. Input is standard HL7/FHIR demographic and clinical fields. Output is a deduplicated Master Patient Index—a drop-in replacement for existing MPI systems. Match and reject thresholds are configurable, allowing operators to tune the

precision–recall trade-off for their clinical context.

Open source and reproducible. The complete pipeline is implemented as a DVC directed acyclic graph with separate training and inference tracks. All stages run in Docker containers. Source code, model, adapter, and dataset are publicly available (see Acknowledgments).

3.5 Limitations and Future Work

Limitations. ClinFuse is evaluated on synthetic data (Synthea), which, while clinically realistic, may not capture all real-world patterns such as identity theft, deliberate data falsification, or extreme data sparsity in safety-net hospitals. The fusion weights are tuned for this dataset and would require calibration on real EHR data. Synthetic evaluation data may differ in difficulty from real-world data, so reported metrics demonstrate the approach's viability rather than guarantee production performance.

Future work. Key directions include: (1) validation on de-identified real EHR data to confirm generalization; (2) *tiered model escalation* where uncertain pairs route from smaller models to larger ones and ultimately to human-in-the-loop review, creating a graduated safety net for the most ambiguous cases; (3) *multimodal resolution* leveraging MedGemma's vision capabilities to incorporate radiology images and scanned documents as additional identity signals; (4) joint demographic–clinical fine-tuning in a single end-to-end model; and (5) embedding-based candidate matching using learned patient representations for fast approximate nearest-neighbor search, replacing hand-crafted blocking rules to scale to millions of records.

4. Conclusion

This work demonstrates that MedGemma—pretrained for clinical QA and medical imaging, never designed for entity resolution—can be effectively repurposed for patient identity matching through QLoRA adaptation. Achieving 0.93 F1 on this novel pairwise classification task, its integration into the ClinFuse pipeline lifts end-to-end recall from 0.991 to a perfect 1.000 while maintaining 99.5% precision. The system runs entirely on-premises on a single consumer GPU, requires no cloud API, and invokes the LLM only for the small fraction of pairs that probabilistic matching cannot resolve. These properties—high accuracy, full privacy, low cost, and interpretable fusion—make MedGemma-powered entity resolution a practical path toward the 99.9% matching accuracy that the MATCH IT Act demands and that patient safety requires.

Entity resolution represents a genuinely novel application of MedGemma—repurposing medical language understanding for a task entirely outside the model's pretraining distribution. The success of this transfer suggests that medical foundation models have broader utility than their original design scope implies: clinical NLP capabilities can power infrastructure tasks that underpin the entire healthcare data ecosystem. As health information exchange scales under TEFCA, MedGemma's combination of clinical reasoning, edge deployability, and adaptability through parameter-efficient fine-tuning offers a foundation for safe, accurate, and privacy-preserving patient matching at national scale.

Acknowledgments

This work uses MedGemma [12] (Google Health AI) as the foundation model and Splink [13] for probabilistic linkage. Synthetic data generated with Synthea [14]. Developed for the Kaggle MedGemma Impact Challenge [19].

Resources:
Code: https://github.com/abicyclerider/clinfuse
Model: https://huggingface.co/abicyclerider/medgemma-4b-entity-resolution-text-only
Dataset: https://huggingface.co/datasets/abicyclerider/entity-resolution-pairs
Video: *[URL TBD before submission]*

References

[1] ECRI Institute PSO, "Deep dive: Patient identification," 2016.

[2] J. Western et al., "SEE ALTERNATE MRN: Duplicate charts in hospitalized patients," *BMJ Qual. Saf.*, 2026.

[3] G. Morris et al., "Patient matching is a challenge," *Perspect. HIM*, 2016.

[4] Black Book Market Research, "Provider interoperability and patient record error rates," 2018.

[5] ONC, "Patient identification and matching final report," 2014.

[6] K. Harron et al., "Linkage error bias," *Int. J. Epidemiol.*, 48(6):2050–2060, 2019.

[7] U.S. Congress, "MATCH IT Act," H.R. 2002, 2025.

[8] ONC, "TEFCA v2.1," 2024.

[9] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *JASA*, 64(328):1183–1210, 1969.

[10] Z. Wu et al., "MedLink: De-identified patient record linkage," *Proc. KDD*, 2023.

[11] B. Qian et al., "PRISM: Patient records interpretation for clinical trial matching using LLMs," *npj Digit. Med.*, 7, 2024.

[12] Google Health AI, "MedGemma model card," 2025.

[13] R. Linacre et al., "Splink: Probabilistic record linkage at scale," *Int. J. Pop. Data Sci.*, 2022.

[14] J. Walonoski et al., "Synthea: Generating synthetic patients and EHRs," *JAMIA*, 25(3):230–238, 2018.

[15] T. Dettmers et al., "QLoRA: Efficient finetuning of quantized LLMs," *NeurIPS*, 2023.

[16] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[17] S. Gao et al., "Fine-tuned small LLMs vs. zero-shot large LLMs for classification," *arXiv:2406.08660*, 2024.

[18] S. Singh et al., "Classification heads: 130x faster than prompting," *arXiv:2411.05045*, 2024.

[19] F. Mahvar et al., "MedGemma Impact Challenge," Kaggle, 2026.

[20] R. Peeters and C. Bizer, "Entity matching using large language models," *arXiv:2310.11244*, 2023.

[21] Y. Li et al., "On leveraging large language models for enhancing entity resolution: A cost-efficient approach," *arXiv:2401.03426*, 2024.

[22] T. Wang et al., "Match, compare, or select? An investigation of large language models for entity matching," *Proc. COLING*, 2025.

[23] C. Cao et al., "Linking patient records at scale with a hybrid approach combining contrastive learning and deterministic rules," *medRxiv*, 2025.