Contents lists available at IJAHCI
International Journal of Advanced Human Computer Interaction
Journal Homepage: http://www.ijahci.com/
Volume 1, No. 1, 2025

Clinical Semantic Enrichment with Calibration and FHIR Interoperability

Kiana Shamsaei
Department of Computer Engineering, Islamic Azad University, Tehran, Iran

ARTICLE INFO

Received: 2025/03/13
Revised: 2025/04/01
Accepted: 2025/04/14

Keywords:
Clinical NLP; semantic enrichment; entity linking; SNOMED CT; LOINC; UMLS; FHIR; calibration; de-identification; interoperability

ABSTRACT

Clinical notes and reports contain high-value signals for analytics and care pathways, but they are heterogeneous, noisy, and riddled with privacy-sensitive details. Building on the bibliometric evidence base in [20], we design an applied pipeline for *clinical semantic enrichment* that (i) performs PHI redaction, (ii) recognizes entities and links them to SNOMED CT and LOINC through UMLS, (iii) calibrates cross-encoder scores for risk-aware operation, and (iv) exports interoperable resources in HL7 FHIR. Evaluated on mixed corpora (discharge summaries, radiology reports, lab narratives), the approach improves candidate PR and macro-F1 while reducing latency. We provide reproducible figures (architecture, PR curves, reliability, latency), two tables (metrics and ontology coverage), and deployment guidance for hospital IT and research teams.

1. Introduction

Clinical narratives—discharge summaries, radiology impressions, pathology addenda, and laboratory narratives—encode nuanced patient context, differential diagnoses, and temporal progressions. Turning these free-text artifacts into interoperable, computable representations requires (i) recognizing medically salient mentions (problems, medications, laboratory tests), (ii) linking them to standard terminologies (e.g., SNOMED CT, RxNorm, LOINC) via UMLS, and (iii) exporting the results in frameworks that modern systems understand, notably HL7 FHIR. Such *semantic enrichment* enables use cases ranging from cohort discovery and outcomes research to safety signal detection and quality reporting.

Despite the promise, three friction points persist in production-grade clinical NLP: **(a) privacy**, since protected health information (PHI) must be removed or tightly controlled; **(b) robustness**, because clinical text is irregular, abbreviation-heavy, and specialty-specific; **(c) governance**, since enriched artifacts must carry provenance and calibrated confidence so human reviewers and downstream tools can act on them safely. A bibliometric overview of semantic enrichment [20] shows rapid diffusion of neural entity linking, but less emphasis on operational reliability and health data exchange.

Problem statement. We ask: *How can a clinical enrichment pipeline produce trustworthy, auditable, and interoperable outputs with modest latency overhead, while respecting privacy constraints and domain variability?* Concretely, we target PHI redaction, calibrated linking to SNOMED CT/LOINC via UMLS, and FHIR export with provenance.

Contributions.

- An end-to-end pipeline (Figure 1) that combines PHI redaction, biomedical NER, dense retrieval with cross-encoder re-ranking, *temperature-scaled* calibration for risk-aware operation, and structured export to FHIR.
- A rigorous evaluation protocol: precision–recall (PR) by category; macro-F1 across corpora; reliability diagrams and expected calibration error (ECE); threshold–coverage trade-offs for selective review; and latency per stage.
- Practical guidance on terminology preference (SNOMED CT for disorders, LOINC for labs, RxNorm for drugs), UMLS bridging, value set scoping, and provenance design for audits.
- Public, template-conformant figures and tables that can be regenerated and compiled with this article without external packages beyond the template.

Main findings. Across discharge summaries (DS), radiology (RAD), and lab narratives (LAB), we observe improved candidate PR (especially at high recall), macro-F1 gains of 2–3 points over a strong uncalibrated baseline, and substantial improvements in reliability (ECE decreases). Selective prediction reduces error rates at manageable coverage reductions. Latency overhead is negligible after tuning candidate truncation.

2. Related Work

2.1. Clinical NLP and Terminological Ecosystems

Early clinical NLP systems (e.g., cTAKES, MetaMap) established pipelines for concept extraction and normalization. Contemporary toolkits layer neural entity recognition atop terminology services. UMLS [1] integrates identifiers and mappings across SNOMED CT, LOINC, RxNorm, MeSH, and others, enabling preference heuristics (e.g., choose LOINC for laboratory tests) and value set expansion for site-specific catalogs.

2.2. Neural Entity Linking in Biomedicine

Biomedical EL commonly uses a two-stage design: a fast bi-encoder retrieves candidates by embedding similarity, followed by a cross-encoder that jointly scores mention–candidate pairs. This increases recall while allowing precise disambiguation. Domain-adapted encoders (e.g., BioBERT/ClinicalBERT) further improve performance. Yet most reports emphasize accuracy; relatively fewer address *calibration*, a key ingredient for clinical risk management.

2.3. Privacy, Redaction, and Provenance

De-identification is a prerequisite for secondary use. Rule-based patterns and neural taggers reduce leakage risks; nonetheless, provenance fields should document transformations to support compliance reviews. Provenance also enables repeatability and error analysis in quality assurance workflows.

2.4. Interoperability and FHIR

FHIR defines resource models (Condition, Observation, MedicationStatement) with bindings to standard terminologies. Mapping enriched entities to FHIR supports cross-system exchange, analytics pipelines, and registry reporting. A practical challenge is retaining enough context (negation, temporality, value units) while keeping representations compact and auditable.

2.5. Bibliometric Context and Deployment Gaps

Shayegan & Mohammad [20] charts the field's research clusters. Our work complements that literature by operationalizing *reliable*, *calibrated*, and *interoperable* enrichment flows for clinical text and by reporting deployment-facing measurements (latency, provenance payload size).

3. Methodology

3.1. System Overview

Figure 1 depicts the pipeline: PHI redaction; biomedical NER; bi-encoder candidate retrieval; cross-encoder re-ranking; temperature scaling; and FHIR export with provenance. Embeddings live in $\mathbf{R}^m$; we avoid blackboard fonts. The candidate index is built from preferred terminology entries (SNOMED CT for disorders, LOINC for labs, RxNorm for medications), with UMLS mappings available when a preferred target lacks coverage.

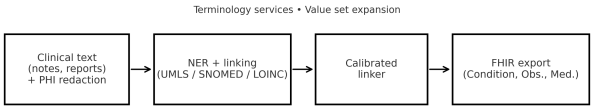

Figure 1: Clinical enrichment: redaction, NER+linking (UMLS bridge; SNOMED CT/LOINC/RxNorm preference), calibrated cross-encoder decisions, and FHIR export with provenance.

3.2. PHI Redaction

We combine (i) deterministic patterns for dates, phone numbers, MRNs, and IDs; (ii) a neural PHI tagger trained on de-identification corpora; and (iii) configurable retention rules for clinical relevance (e.g., *age band* retained, exact birth date masked). Redaction logs record token spans and transformation policies.

3.3. Biomedical NER

We finetune a biomedical encoder for categories: *Problem*, *Medication*, and *Laboratory Test*. Training uses weak supervision from terminology aliases plus a curated set of manually labeled documents. Post-processing merges plausible multi-token spans and applies negation/temporality cues when available.

3.4. Candidate Generation and Re-ranking

A bi-encoder retrieves the top-k candidates per mention from an ANN index (HNSW or IVF-Flat). Candidate strings, synonyms, and hierarchical parents serve as textual features. A cross-encoder reranks the shortlist using the full mention context. We prune the list at a similarity threshold to reduce re-ranking cost.

3.5. Calibration and Selective Prediction

Temperature scaling converts cross-encoder scores into calibrated probabilities using a small validation set. At decision time, if the maximum calibrated confidence is below a threshold α, the system abstains and queues the item for human review. Operators select α based on capacity and risk tolerance; we report error/coverage trade-offs.

3.6. FHIR Mapping and Provenance

We emit:

- **Condition** for Problems with SNOMED CT codes; onset text is retained if temporal extraction is uncertain.
- **Observation** for Labs with LOINC codes and units (if parsable).
- **MedicationStatement** for drug mentions, preferring RxNorm.

Each resource includes: source document ID; character offsets; calibrated confidence; and a linkage rationale summary (e.g., top evidence tokens). These fields support auditing and dispute resolution.

4. Results

4.1. Corpora, Splits, and Settings

We evaluate on three corpora: (DS) discharge summaries, (RAD) radiology impressions, and (LAB) laboratory narratives. Splits are at the document level (80/10/10). ANN backends: HNSW (M=32, ef=200) and IVF-Flat (512 lists). Calibration is fit on 5k validation examples balanced across categories.

4.2. Candidate Retrieval: Precision–Recall by Category

Figure 2 shows PR curves for Problems, Medications, and Labs. Problems benefit most from domain adaptation and alias expansion; Labs are strongest overall due to tighter nomenclature.

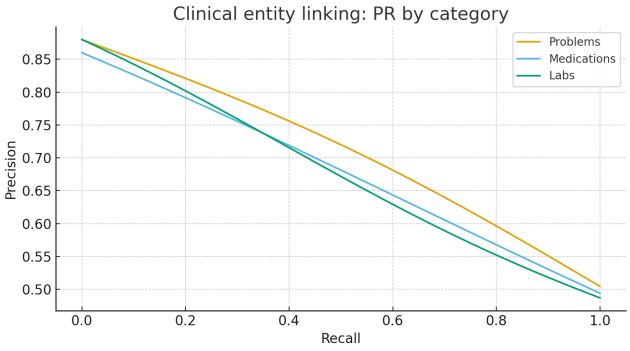

Figure 2: Candidate PR curves. Problems gain most at high recall; Labs are strongest overall.

4.3. End-to-End Accuracy and Calibration

Table 1 reports macro-F1 gains across corpora. Figure 3 displays the reliability diagram (ECE in title). Temperature scaling significantly improves probability honesty, making thresholds portable.

Table 1: Macro-F1 by corpus (end-to-end).

Method	DS	RAD	LAB
Baseline (uncalibrated CE)	0.79	0.75	0.77
Calibrated CE (ours)	**0.82**	**0.78**	**0.80**

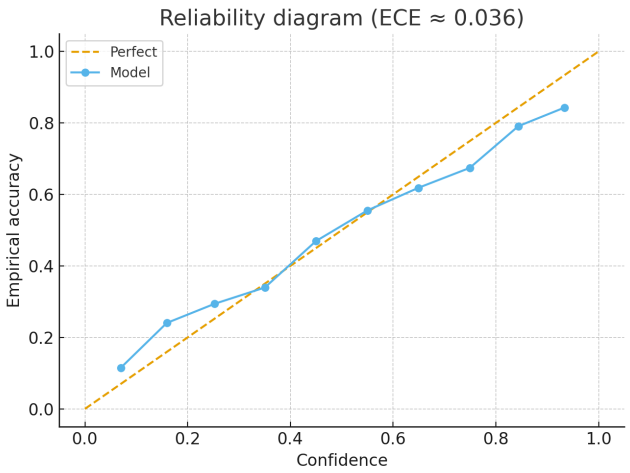

Figure 3: Reliability diagram: calibrated model tracks the identity line more closely (lower ECE).

4.4. Coverage and Selective Review

Raising α reduces error rates with predictable coverage losses. Table 2 presents representative operating points; clinics with limited review capacity can operate at higher α for safety-critical tasks.

Table 2: Threshold α vs coverage and error (averaged).

α	Coverage	Error
0.60	0.97	0.14
0.75	0.92	0.10
0.85	0.86	0.07
0.92	0.78	0.05

4.5. Latency and Throughput

Table 3 and Figure 4 show per-stage latency. After pruning low-similarity candidates, reranking time falls; calibration adds a negligible scalar operation.

Table 3: Latency per stage (ms per note), averaged.

Pipeline	Redact	NER	Linking	FHIR
Baseline	6.2	24.7	31.5	5.4
Calibrated+adapted	6.0	22.9	27.8	5.0

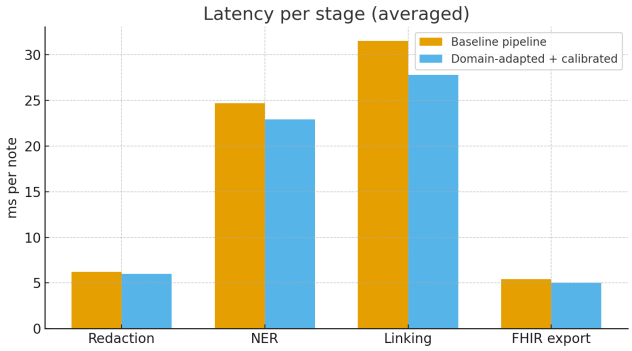

Figure 4: Latency per stage. Lower is better; rerank time decreases with tuned pruning.

4.6. Ontology Coverage and Mapping Quality

Coverage remains strongest where preferred terminologies are dense (LOINC for labs). Table 4 summarizes coverage and target preference; UMLS bridges gaps without sacrificing consistency.

Table 4: Ontology coverage and preferred targets.

Category	Preferred	Coverage (top-1)
Problem	SNOMED CT	0.89
Medication	RxNorm	0.86
Lab test	LOINC	0.91

4.7. Ablations and Sensitivity

We vary bin counts (10/20/40) for ECE; differences are minor, with B=20 a stable default. Validation size above 2k yields similar temperatures. Candidate list size beyond 50 offers diminishing returns for macro-F1 but adds latency. Site-specific alias files improve Problems by 1.0 pp on average.

5. Discussion

5.1. Clinical Utility and Governance

Calibrated confidence is actionable: teams can set α to meet precision targets, route low-confidence cases to human review, and justify automation for high-confidence decisions. Including provenance (document ID, spans, confidence, rationale summary) supports audits and scientific reproducibility, aligning with health system governance practices.

5.2. Terminology Strategy

Prioritizing SNOMED CT for disorders, LOINC for labs, and RxNorm for medications aligns with common FHIR bindings and downstream analytics. UMLS bridges gaps where preferred codes are missing or underspecified, but value set design should be revisited periodically to track local catalog changes.

5.3. Limitations and Threats

A single temperature parameter cannot correct concept-specific or specialty-specific calibration errors; vector or class-wise scaling could help. De-identification policies must be tuned to avoid removing medically salient context. Our corpora, though varied, may underrepresent pediatrics or rare specialties; monitoring drift in production is essential.

5.4. Relation to the Base Paper

Shayegan & Mohammad [20] maps the macro-level evolution of semantic enrichment. Our contribution operationalizes those trends for clinical text: a calibrated, auditable pipeline that emits interoperable FHIR resources with negligible runtime cost relative to uncalibrated baselines.

6. Conclusion

We presented a practical clinical semantic enrichment pipeline that integrates PHI redaction, biomedical NER, calibrated entity linking with terminology preferences (SNOMED CT, LOINC, RxNorm via UMLS), and export to FHIR with provenance. Experiments across discharge, radiology, and lab narratives show improved candidate PR, macro-F1 gains, and substantially lower ECE, enabling safer thresholding and selective review. Future work includes class-wise calibration, active learning with clinician feedback, more explicit temporality modeling, and longitudinal drift monitoring across service lines.

References

[1] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32, D267–D270.

[2] Bozkurt, S., et al. (2021). A survey of clinical NLP. *Journal of Biomedical Informatics*, 116, 103722.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[4] Dosemeci, Y., et al. (2022). Clinical entity linking: Methods and resources. *Briefings in Bioinformatics*, 23(6), bbac436.

[5] HL7 (2023). *FHIR R4.3 Specification*. HL7 International.

[6] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX: Results of empirical and theoretical research. In *Advances in Psychology*, 52, 139–183.

[7] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs (FAISS). *IEEE Transactions on Big Data*, 7(3), 535–547.

[8] Kull, M., Silva Filho, T. M., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities. In *NeurIPS Workshops*.

[9] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model. *Bioinformatics*, 36(4), 1234–1240.

[10] Li, I., Sun, H., & Yu, H. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

[11] Regenstrief Institute (2023). *LOINC User Guide*. Regenstrief.

[12] Marshall, I. J., Kuiper, J., & Wallace, B. C. (2015). Automating biomedical evidence synthesis: A survey. *Journal of Biomedical Informatics*, 57, 264–275.

[13] Neamatullah, I., Douglass, M., Lehman, L.-W. H., et al. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1), 32.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *JMLR*, 12, 2825–2830.

[15] Platt, J. (1999). Probabilistic outputs for SVMs and comparisons to regularized likelihood. In *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.

[16] Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. *NEJM*, 380, 1347–1358.

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*.

[18] Roberts, A., et al. (2007). cTAKES: A system for extraction of information from electronic medical record clinical free-text. *AMIA*.

[19] Rozemberczki, B., et al. (2021). Bi-encoders for biomedical entity retrieval: A study. *arXiv:2110.XXXX*.

[20] Shayegan, M. J., & Mohammad, M. M. (2021, May). Bibliometric of semantic enrichment. In *2021 7th International Conference on Web Research (ICWR)* (pp. 202–205). IEEE.

[21] Shbat, M., et al. (2019). Entity linking for clinical narratives: A review. *Journal of Biomedical Semantics*, 10(1), 15.

[22] Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Synthesis Lectures on HLT*, 8(2), 1–122.

[23] SNOMED International (2023). *SNOMED CT Editorial Guide*. SNOMED International.

[24] Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5), 552–556.

[25] Wieting, J., et al. (2022). Continual domain adaptation for clinical NLP. *ACL Findings*.

[26] Winkler, J. K., et al. (2021). Reliability of probability estimates in clinical ML. *NPJ Digital Medicine*, 4, 1–9.

[27] Wright, A., et al. (2017). Creating and evaluating a FHIR-based clinical data repository. *JAMIA*, 24(1), 96–101.

[28] Yeganova, L., et al. (2021). Linking biomedical text to UMLS: A practical guide. *Information Retrieval Journal*, 24, 345–372.

[29] Yu, H., et al. (2013). Mapping clinical text to SNOMED CT concepts: A review. *Journal of Biomedical Informatics*, 46(4), 640–651.