A Multilingual Workflow for Constructing Semantic Bridge Edges between Local Clinical Knowledge Graphs and Standard Clinical Ontologies

Artid Boonrerng*, Nouh Sabri Elmitwally, Mohamed Ragab, Edlira Vakaj and Hadeel Saadany

Department of Computer Science, Birmingham City University, Birmingham, B4 7BD, United Kingdom

Abstract

Integrating local clinical knowledge graphs with standard clinical ontologies is difficult when local expressions differ in language, lexical form, abbreviation, or conceptual specificity. This paper presents a multilingual workflow for retrieving, reranking, selecting, and materialising candidate correspondences as evidence-bearing SEMANTIC_BRIDGE relations. A materialised bridge records an automatically selected correspondence and does not by itself assert semantic equivalence. Candidate SNOMED CT concepts are generated using BM25 lexical retrieval and LaBSE multilingual semantic retrieval, then reranked using lexical, semantic, contextual, and ontology-based evidence before thresholded selection. The Thai case study uses a purpose-built clinical knowledge graph with 124 nodes and 236 relations. Candidate-union Recall@10 ($K = 10$ per retriever) was 0.742 (72/97 mapped test nodes). Lexical achieved the highest F1 (0.338), while Combined achieved the highest precision (0.541). Complete emitted 36 bridges with precision 0.472, recall 0.175, F1 0.256, and coverage 0.367. Of these, 32/36 were connected to a reviewed target within six undirected active inferred `is-a` edges, although only 17 were exact matches and 15/19 incorrect bridges were also path-connected. Ontology proximity should therefore not be interpreted as semantic equivalence. A Spanish run demonstrates procedural reuse after replacing source-language resources, but identifier provenance prevents a defensible cross-language accuracy comparison.

Keywords

Semantic Bridge Edges, Knowledge Graph Construction, Clinical Knowledge Graphs, Ontology Alignment, Clinical Ontologies, Low-Resource Languages

CEUR-WS.org/Vol-4335/paper7.pdf

1. Introduction

Knowledge graphs support integration, querying, provenance tracking, and reuse across heterogeneous information sources [1]. In clinical informatics, connecting local knowledge to standard terminologies is difficult because equivalent or closely related concepts may be expressed through language-specific terms, abbreviations, synonyms, transliterations, or locally preferred descriptions. Resources such as UMLS illustrate the value of terminology integration [2], while concept normalisation and biomedical entity linking provide mechanisms for mapping textual expressions to standard identifiers [3, 4]. However, ambiguity, conceptual specificity, and cross-language lexical variation remain important challenges [5]. Multilingual representations such as LaBSE can improve candidate retrieval when direct lexical overlap is limited [6, 7].

Established standards already support mapping representation and provenance. SKOS [8] and SSSOM [9] provide mechanisms for expressing mappings and associated metadata, while PROV-O [10] supports provenance. The problem addressed here is therefore complementary: how candidate correspondences can be automatically retrieved, reranked, selected, and materialised as persistent relations within an integrated clinical knowledge graph while retaining the evidence and provenance of the construction decision.

Proceedings of the 5th International Workshop on Natural Language Processing for Knowledge Graph Construction (NLP4KGC 2026), co-located with the 25th International Semantic Web Conference (ISWC 2026), Bari, Italy, October 26, 2026

*Corresponding author.

artid.boonrerng@mail.bcu.ac.uk (A. Boonrerng); nouh.elmitwally@bcu.ac.uk (N. S. Elmitwally); mohamed.ragab@bcu.ac.uk (M. Ragab); edlira.vakaj@bcu.ac.uk (E. Vakaj); hadeel.saadany@bcu.ac.uk (H. Saadany)

0009-0002-4883-3039 (A. Boonrerng)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-19

We propose a multilingual workflow that separates candidate retrieval, evidence-based reranking, thresholded selection or abstention, and persistent graph materialisation. Lexical and multilingual semantic retrieval generate candidate SNOMED CT concepts, which are reranked using lexical, semantic, contextual, and ontology-based evidence. A selected candidate may then be materialised as a SEMANTIC_BRIDGE relation retaining its reranking score, component evidence, provenance, source language, target ontology, and ontology release. The relation records the outcome of the construction procedure and does not by itself assert semantic equivalence. SNOMED CT is used as the target ontology in the Thai case study because it provides both standardised clinical concepts and an explicit hierarchy [11]; a separate Spanish run examines procedural transferability after replacing source-language resources.

The paper makes three contributions: (1) a multilingual workflow that explicitly separates retrieval, reranking, selection, and persistent materialisation; (2) a controlled Thai–SNOMED CT evaluation separating candidate retrieval, exact bridge correctness, coverage, and ontology connectivity; and (3) a procedural-transferability study using Spanish source resources without claiming equivalent cross-language mapping effectiveness. The study addresses:
RQ1: How do lexical, semantic, contextual, and ontology-based evidence configurations affect candidate selection and semantic bridge construction?
RQ2: To what extent do emitted bridges achieve exact reference agreement and structural connectivity in the target ontology?
RQ3: Can the language-independent construction workflow be reused after replacing source-language resources?

2. Related Work

Biomedical entity linking and concept normalisation associate textual mentions with concepts in controlled terminologies using lexical, statistical, neural, or hybrid approaches [3, 4]. A persistent difficulty is ambiguity: a local expression may correspond to a broader, narrower, related, or non-equivalent ontology concept [5]. Multilingual settings introduce additional lexical variation. Portelli et al. [6] studied multilingual medical term normalisation, while LaBSE provides a shared multilingual embedding space suitable for semantic candidate retrieval across languages [7].

Ontology-alignment systems increasingly combine heterogeneous evidence for correspondence discovery. OLaLa [12] applies large language models to ontology matching, while OntoAligner [13] combines multiple alignment strategies within a modular framework. These approaches motivate the use of lexical, semantic, and structural evidence, but correspondence representation itself is already supported by established standards.

SKOS defines mapping predicates including `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, and `skos:relatedMatch` [8]. SSSOM supports mappings together with predicates, justification, confidence, provenance, and source–target metadata [9], while PROV-O provides a general provenance vocabulary [10]. SNOMED CT similarly provides mapping reference sets for representing mappings to other coding and terminology systems [14]. SeMRA further demonstrates the value of retaining biomedical mappings as graph-based, provenance-aware resources that support inspection and reasoning [15].

The present study therefore does not propose a replacement mapping standard. Instead, it addresses the operational process by which multilingual candidate correspondences are retrieved, reranked using heterogeneous evidence, selected under an explicit decision rule, and materialised as persistent graph relations. SEMANTIC_BRIDGE records the outcome and supporting evidence of this construction process but does not by itself assert semantic equivalence. This complements recent NLP4KGC work on ontology learning and semi-automatic knowledge graph construction [16, 17]: here, both the local graph and target ontology already exist, and the task is to connect them rather than construct either resource from scratch.

3. Proposed Method

Figure 1 summarises the proposed workflow. Language-specific preprocessing is followed by lexical and multilingual semantic candidate retrieval. The retrieved candidate sets are merged and subsequently reranked using multiple evidence signals. A thresholded selection step then determines whether a candidate should be selected for materialisation, after which the selected correspondence is materialised as a persistent graph relation. After preprocessing, the workflow separates five stages: candidate retrieval, candidate union, evidence-based reranking, selection or abstention, and graph materialisation.

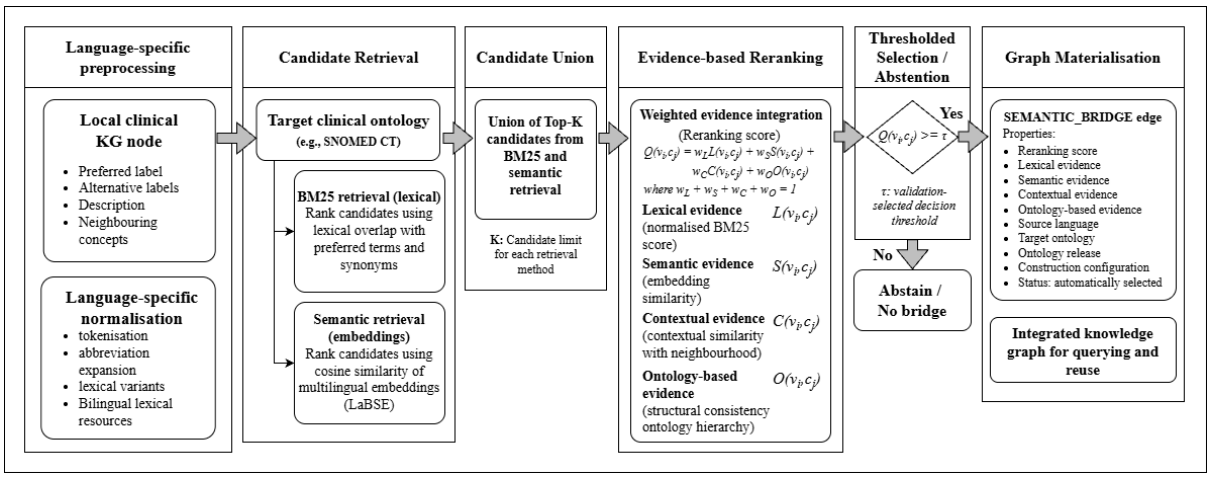


Figure 1: Semantic bridge construction workflow. After language-specific preprocessing, lexical and multilingual semantic candidates are retrieved and merged, reranked using multiple evidence signals, subjected to thresholded selection or abstention, and materialised as persistent graph relations.

3.1. Problem Formulation and Candidate Generation

Let $G_L = (V_L, E_L)$ denote a local clinical knowledge graph and $G_T = (V_T, E_T)$ a target clinical ontology. For a local node $v_i \in V_L$, the workflow seeks candidate target concepts $c_j \in V_T$ and determines whether one of them should be materialised as

$$b_{ij} = (v_i, \text{SEMANTIC_BRIDGE}, c_j). \tag{1}$$

A SEMANTIC_BRIDGE records a correspondence selected by the construction procedure under a particular configuration and target-ontology release. Its existence does not by itself assert that v_i and c_j are semantically equivalent, nor does it assign a specific mapping predicate such as `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, or `skos:relatedMatch`; such semantics would need to be established separately. In the present evaluation, exact agreement with the reviewed SNOMED CT target is used as an evaluation criterion rather than as an intrinsic semantic property of the bridge relation.

This formulation separates candidate discovery from later decision stages. A target concept may be retrieved but subsequently ranked below another candidate or rejected by the selection rule. Conversely, if the reviewed target is absent from the retrieved candidate set, downstream reranking and selection cannot recover an exact bridge. These cases are therefore treated separately as retrieval and selection failures.

Candidate generation uses two complementary retrieval mechanisms. Lexical retrieval uses BM25 [18] with Unicode-aware word tokenisation, $k_1 = 1.5$, and $b = 0.75$. Semantic retrieval uses LaBSE [7] through TensorFlow Hub to encode source and target expressions into 768-dimensional multilingual representations, with candidates ranked by cosine similarity. This multilingual representation allows semantic candidate retrieval across source and target languages without requiring lexical overlap.

For the Thai case study, the lexical query document combines the normalised local expression with available English lexical variants. These variants are obtained from explicitly provided English forms and the Thai–English bilingual lexicon when a corresponding local expression is present. In contrast, LaBSE semantic retrieval encodes the normalised Thai source expression directly against the English target descriptions in the shared multilingual embedding space.

Each retriever returns its top $K = 10$ candidates. For local node v_i, the lexical and semantic candidate sets are denoted by C_i^{lex} and C_i^{sem}, respectively. The candidate pool passed to the reranking stage is their deduplicated union,

$$C_i = C_i^{\text{lex}} \cup C_i^{\text{sem}}. \tag{2}$$

The union therefore contains at most 20 unique candidates for each local node. In the Thai held-out test split, the observed mean was 17.29 candidates per node. Candidate generation is treated strictly as an upstream retrieval stage; no retrieved candidate becomes a persistent bridge until it has passed the subsequent reranking and selection stages.

3.2. Evidence-based Reranking, Selection, and Bridge Materialisation

For each candidate pair (v_i, c_j), the reranking stage uses up to four evidence signals: lexical $L(v_i, c_j)$, semantic $S(v_i, c_j)$, contextual $C(v_i, c_j)$, and ontology-based $O(v_i, c_j)$. These signals support candidate ordering and selection and are not interpreted as calibrated probabilities of mapping correctness.

Lexical and semantic evidence. Lexical evidence is the BM25 score normalised independently for each local node by its maximum non-zero BM25 score:

$$L(v_i, c_j) = \frac{\text{BM25}(v_i, c_j)}{\max_{c \in V_T} \text{BM25}(v_i, c)}. \tag{3}$$

Candidates without a lexical match receive $L(v_i, c_j) = 0$. Semantic evidence is the cosine similarity between L2-normalised LaBSE representations:

$$S(v_i, c_j) = \cos(\mathbf{e}(v_i), \mathbf{e}(c_j)) \,. \tag{4}$$

Although the two signals are numerically bounded, they arise from different scoring procedures and are not assigned the same probabilistic interpretation.

Contextual evidence. For v_i, the context representation combines its description and preferred label with preferred labels from up to ten one-hop local neighbours. The target representation similarly incorporates the target description and preferred terms from up to ten adjacent SNOMED CT concepts. Their LaBSE cosine similarity defines

$$C(v_i, c_j) = \cos(\mathbf{e}_{\text{ctx}}(v_i), \mathbf{e}_{\text{ctx}}(c_j)) \,. \tag{5}$$

The ten-neighbour limit is a predefined heuristic cap. When more than ten neighbours are available, the implementation retains up to the first ten in adjacency iteration order; no relation-aware or relevance-based ranking is applied.

Ontology-based evidence. Each neighbouring local node u is first assigned a provisional target

$$\hat{c}^u = \arg\max_{c \in C_u} [0.5L(u, c) + 0.5S(u, c)] \,. \tag{6}$$

Let $N'(v_i)$ denote neighbouring local nodes with provisional targets, and let $d_T(c_j, \hat{c}^u)$ be the undirected shortest-path distance over active inferred SNOMED CT `is-a` edges. Ontology-based evidence is

$$O(v_i, c_j) = \frac{1}{|N'(v_i)|} \sum_{u \in N'(v_i)} \exp\left(-\frac{d_T(c_j, \hat{c}^u)}{2}\right). \tag{7}$$

Shortest-path search is bounded at six hops; unavailable paths contribute zero, and $O(v_i, c_j) = 0$ when $N'(v_i) = \emptyset$. This is a heuristic structural-compatibility signal rather than a measure of semantic equivalence. Local adjacency is also treated in a relation-agnostic manner.

Reranking and selection. The four evidence signals are combined as

$$Q(v_i, c_j) = w_L L(v_i, c_j) + w_S S(v_i, c_j) + w_C C(v_i, c_j) + w_O O(v_i, c_j), \tag{8}$$

where active weights sum to one. Lexical and Semantic use a single evidence source, while Combined uses $(0.5, 0.5, 0, 0)$. For Complete, weights were selected on the predefined validation split from a grid with 0.1 increments, minimum component weight 0.1, and weights constrained to sum to one. Thresholds were evaluated at 19 score quantiles spanning 0.05–0.95. Configurations were ranked by validation F1, then precision, then the lower threshold. This selected $(0.4, 0.4, 0.1, 0.1)$ for Complete.

For each local node,

$$\hat{c}_i = \arg\max_{c_j \in C_i} Q(v_i, c_j), \tag{9}$$

When candidates receive identical reranking scores, ties are resolved deterministically by ascending target concept identifier. A bridge is then created only when the selected score reaches the configuration-specific threshold:

$$\hat{b}_i = \begin{cases} (v_i, \text{SEMANTIC_BRIDGE}, \hat{c}_i), & Q(v_i, \hat{c}_i) \geq \tau, \\ \varnothing, & Q(v_i, \hat{c}_i) < \tau. \end{cases} \tag{10}$$

The resulting thresholds were 0.670938 for Semantic, 0.820843 for Combined, and 0.767326 for Complete; Lexical used $\tau = 0$. Thresholds were fixed before held-out test evaluation.

Bridge materialisation. Selected correspondences are materialised as explicit SEMANTIC_BRIDGE relations retaining the reranking score, component evidence, source language, target ontology, ontology release, construction configuration, and construction status. The relation records the construction decision and does not by itself assign a mapping predicate such as `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, or `skos:relatedMatch`. Language-specific preprocessing resources may change across languages, while candidate union, reranking, selection, metadata, and graph materialisation remain unchanged.

4. Experimental Setup

4.1. Thai Data, Reference Set, and Target Ontology

The main experiment used a purpose-built Thai clinical knowledge graph containing 124 local clinical nodes and 236 stored relations. The graph was designed to represent local clinical expressions rather than direct copies of English ontology labels and includes formal terminology and local variants across findings, disorders, procedures, medications, and observable concepts. Each node contains a preferred label and, where available, alternative labels, a short description, and English lexical variants.

The 236 stored relations comprised 112 `local_synonym_of`, 104 `related_clinical_concept`, 10 `related_to`, three `investigated_by`, three `treated_by`, two `measured_by`, and two `has_finding` relations. These relation labels were retained in the source graph, but the current contextual and ontology-based evidence components use graph adjacency in a relation-agnostic manner and do not distinguish among relation types during scoring.

The reference mappings connect the Thai local nodes to active SNOMED CT concepts. Candidate mappings were initially generated from terminology retrieval outputs and subsequently reviewed through an AI-assisted terminology review process followed by author verification. The reviewed mappings were recorded and inspected before being used for evaluation. Because they were not independently adjudicated by multiple clinical terminology experts, they are treated as an author-verified silver reference set produced with AI-assisted terminology review rather than as a clinical gold standard. Of the 124 local nodes, 123 were assigned acceptable SNOMED CT targets and one was intentionally retained as a no-match case.

The mapping, no-match, and data-split annotations were stored separately from the local-node records and joined by local identifier before model execution. A predefined split assigned 26 nodes to validation and 98 to held-out testing. The test set contained 97 mapped nodes and one intentional no-match node. Validation nodes were used for configuration selection and were excluded from final held-out reporting.

For the Thai experiment, language-specific preprocessing differs between lexical and semantic retrieval. The semantic source document combines the normalised Thai preferred label, alternative labels, and description and is encoded directly with LaBSE against English SNOMED CT target descriptions in the shared multilingual embedding space. No machine translation is applied to this semantic retrieval path. For BM25 retrieval, the lexical source document additionally incorporates explicitly provided English variants and entries from the Thai–English bilingual lexicon when a corresponding source expression is available. This provides English lexical cues for lexical retrieval while retaining the original Thai expression for multilingual semantic retrieval.

The target ontology was the SNOMED CT International RF2 Snapshot dated 1 May 2026 [19]. The evaluated target space contained 381,251 active concepts. The English retrieval document for each concept combined its preferred term, fully specified name, and up to ten active synonyms. The active inferred `is-a` hierarchy contained 633,868 unique relations after retaining relationships whose source and destination concepts were active. SNOMED CT is used as the case-study target ontology rather than as a requirement of the bridge representation; its concept inventory and explicit hierarchy support both terminology grounding and post-construction graph traversal. An accompanying public repository provides the paper-aligned Colab implementation and a synthetic example dataset; licensed SNOMED CT content and the research dataset are not redistributed.[1]

4.2. Evaluation Protocol

Candidate retrieval was evaluated using Recall@1, Recall@5, and Recall@10 over the 97 mapped held-out test nodes. A reviewed target was counted as recovered at rank k when an acceptable target occurred within either the lexical Top-k or semantic Top-k list, thereby evaluating retrieval independently of subsequent reranking and selection.

Bridge construction was evaluated using precision, recall, F1, and coverage. Precision measures the proportion of emitted bridges that exactly match an acceptable reviewed target; recall uses the 97 mapped test nodes as its denominator, whereas coverage uses all 98 held-out nodes, including the intentional no-match case. For a mapped node, an incorrect emitted target contributes a false positive and leaves the reviewed target unrecovered, thereby also contributing a false negative. An emission for the intentional no-match case contributes a false positive only.

Configuration selection used only the 26-node validation split. Lexical and Semantic used their respective single evidence sources, Combined used the predefined weights (0.5, 0.5, 0, 0), and Complete weights were selected on the validation split. Decision thresholds for Semantic, Combined, and Complete were selected using validation data, while Lexical used the fixed setting $\tau = 0$. All settings were fixed before held-out evaluation (Table 1). The final Complete bridge set was additionally audited against the active inferred SNOMED CT `is-a` hierarchy. For this post-hoc analysis, the hierarchy was treated as undirected and shortest-path connectivity was evaluated up to six edges. This measures structural

[1] https://github.com/artidbcu/nlp4kgc-semantic-bridge-workflow/releases/tag/v1.0.0

proximity rather than semantic equivalence. Uncertainty for Complete was estimated using 1,000 node-level bootstrap resamples with random seed 42.

4.3. Spanish Procedural Transferability

A separate Spanish run examined whether the construction procedure could be reused after replacing the source-language resources. The controlled source resource was constructed from active Spanish-language SNOMED CT descriptions whose concept identifiers were also present in the International target release. Using random seed 42, 250 source concepts were sampled, comprising 50 validation nodes and 200 test nodes. Each source node contained a Spanish preferred term and available alternative descriptions.

Table 1
Principal data and experimental settings for the Thai evaluation.

Property	Value
Local nodes / relations	124 / 236
Validation / test nodes	26 / 98
Mapped / no-match test nodes	97 / 1
SNOMED CT release	1 May 2026
Active concepts / inferred `is-a` relations	381,251 / 633,868
BM25 parameters	$k_1 = 1.5,\ b = 0.75$
Candidate cut-off per retriever	$K = 10$
Mean / maximum merged candidates	17.29 / 20
Combined weights	$(0.5, 0.5, 0, 0)$
Complete weights	$(0.4, 0.4, 0.1, 0.1)$
Decision thresholds (Sem./Comb./Comp.)	0.670938 / 0.820843 / 0.767326
Maximum ontology distance	6 hops
Bootstrap resamples / seed	1,000 / 42

A source-side graph was constructed from lexical similarity among the Spanish source descriptions. Character 3–5-gram TF–IDF cosine similarity was used to identify neighbouring source nodes, with up to two neighbours retained per node and represented using `source_text_related` relations. No English lexical variants or Spanish–English bilingual lexicon were supplied for this run. LaBSE therefore provided the multilingual semantic retrieval path, while lexical BM25 operated directly on Spanish source expressions against the English target index.

The Spanish source graph should therefore be interpreted as a controlled procedural resource rather than as an independently developed local clinical knowledge graph. The purpose of this experiment is to determine whether the same candidate-union, evidence, thresholded-selection, metadata, and graph-materialisation procedure can be executed after changing the source-language resources. It is not intended to demonstrate general Spanish clinical terminology-mapping effectiveness.

The recorded run processed all 250 source nodes and produced outputs for the 200-node test partition. A separate 64-node pilot reference resource containing 124 relations was also available; however, its local identifier namespace could not be deterministically reconciled with that of the final predictions. The two resources were therefore not combined for quantitative evaluation. Accordingly, no independent reference-based Spanish precision, recall, F1, or cross-language accuracy comparison is reported.

5. Results and Discussion

5.1. Candidate Retrieval and Configuration Operating Points

Each retriever returned up to $K = 10$ candidates. Candidate-union Recall@1 was 0.443 (43/97), increasing to 0.649 (63/97) at rank 5 and 0.742 (72/97) at rank 10. Twenty-five of the 97 mapped held-out targets were therefore absent from the union of the two Top-10 candidate lists. Because subsequent

reranking and selection operate only on retrieved candidates, these cases place a direct upper constraint on exact bridge recovery.

Figure 2 compares the held-out operating-point metrics of the four evidence configurations. The Lexical configuration emitted a bridge for all 98 test nodes and achieved precision 0.337, recall 0.340, F1 0.338, and coverage 1.000. Semantic emitted 46 bridges, giving precision 0.500, recall 0.237, F1 0.322, and coverage 0.469. Combined emitted 37 bridges and achieved the highest precision, 0.541, with recall 0.206, F1 0.299, and coverage 0.378. Complete emitted 36 bridges, of which 17 exactly matched an acceptable reviewed target, giving precision 0.472, recall 0.175, F1 0.256, and coverage 0.367.

At these validation-selected operating points, adding contextual and ontology-based evidence did not improve conventional F1 over the simpler configurations. Instead, the results show a coverage–selectivity trade-off. Lexical maximised coverage but also emitted 65 bridges that did not match an acceptable reviewed target, whereas the thresholded configurations emitted substantially fewer relations.

These comparisons should not be interpreted as a matched-coverage ablation of the individual evidence components. Each configuration uses its own validation-selected operating point and consequently emits a different number of bridges. Differences in precision, recall, and F1 therefore reflect both candidate scoring and the resulting level of abstention. The results support comparison of the complete end-to-end configurations, but they do not establish the isolated marginal effect of an evidence component at equal coverage.

The central contribution therefore does not depend on the Complete configuration outperforming every baseline. Rather, the workflow makes the transition from retrieved candidate to persistent graph relation an explicit decision and retains the evidence and provenance required to inspect that decision.

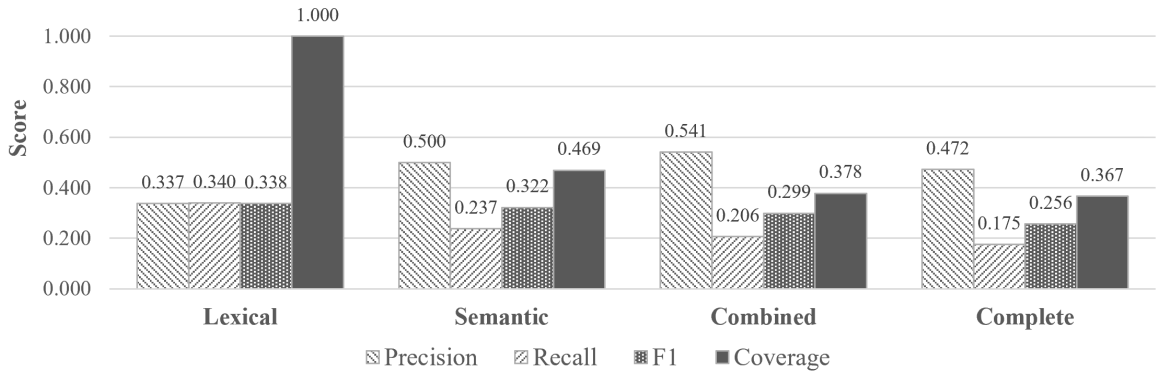


Figure 2: Held-out operating-point metrics for the four semantic bridge construction configurations. Precision is calculated over emitted bridges, recall over the 97 mapped test nodes, and coverage over all 98 test nodes, including the intentional no-match case; F1 is the harmonic mean of precision and recall. Because the configurations use different validation-selected operating points, the comparison is not a matched-coverage ablation.

5.2. Complete Bridge Set, Uncertainty, and Ontology Connectivity

The Complete configuration emitted 36 bridges over the 98 held-out nodes, giving coverage 0.367. Seventeen bridges exactly matched an acceptable reviewed target, while 19 selected a different concept. The remaining 62 nodes received no Complete bridge. The intentional no-match case received candidates during retrieval but no final bridge. Because only one such case exists, this observation is treated as a diagnostic example rather than evidence of general open-world abstention performance.

Table 2 reports bootstrap uncertainty and the post-hoc ontology-connectivity audit. The Complete configuration achieved precision 0.472 (95% CI [0.305, 0.629]), recall 0.175 ([0.102, 0.247]), F1 0.256 ([0.152, 0.350]), and coverage 0.367 ([0.275, 0.459]).

Thirty-two of the 36 emitted bridges were connected to an acceptable reviewed target within at most six undirected active inferred `is-a` edges. This aggregate proportion requires careful interpretation because all 17 exact matches necessarily have ontology distance zero. The incorrect bridges therefore provide a more informative internal comparator: 15 of the 19 incorrect bridges (78.95%) were also connected within the six-hop limit.

Among these 15 incorrect but connected bridges, seven were one hop from an acceptable reviewed target, six were two hops away, one was three hops away, and one was four hops away. Thus, 13/15 were within only one or two hierarchy edges despite being incorrect under exact reference evaluation. Ontology proximity is therefore a weak discriminator of mapping correctness in this experiment rather than an alternative correctness criterion.

The case `TH_035_1` illustrates this distinction. Its reviewed target was successfully retrieved, but the Complete configuration selected a different SNOMED CT concept only one hierarchy edge away. The emitted bridge was therefore structurally close while remaining incorrect under exact reference evaluation.

Table 2
Complete configuration results, bootstrap uncertainty, and ontology-connectivity audit.

Measure	Result
Emitted bridges	36
Exact reviewed-target matches	17/36 (47.22%)
Precision, 95% CI	0.472 [0.305, 0.629]
Recall, 95% CI	0.175 [0.102, 0.247]
F1, 95% CI	0.256 [0.152, 0.350]
Coverage, 95% CI	0.367 [0.275, 0.459]
Connected within ≤ 6 undirected `is-a` edges	32/36 (88.89%)
Incorrect bridges path-connected	15/19 (78.95%)
Incorrect connected distances (1/2/3/4 hops)	7 / 6 / 1 / 1

Ontology connectivity is consequently treated as a complementary graph property. It can indicate that an emitted bridge provides access to the surrounding target-ontology structure, but it cannot determine whether the selected target is semantically equivalent, broader, narrower, or otherwise inappropriate for the local concept. Mapping correctness and graph connectivity should therefore remain separate evaluation dimensions.

5.3. Qualitative Failure Modes

Table 3 illustrates representative outcomes at different stages of the construction workflow. The examples distinguish failure to retrieve an acceptable target from failure to select it after successful retrieval and from threshold-driven abstention.

Table 3
Representative retrieval and construction outcomes.

Case	Interpretation
`TH_042_2`	Reviewed target retrieved and exact bridge emitted: successful retrieval and selection.
`TH_035_1`	Reviewed target retrieved but a different target emitted: selection failure despite successful retrieval.
`TH_002_1`	Reviewed target retrieved but no bridge emitted: abstention after successful retrieval.
`TH_001_1`	Reviewed target absent from the union of the two Top-10 candidate lists and no bridge emitted: retrieval failure.
`TH_NM_004`	Candidates retrieved but no bridge emitted for the intentional no-match case: diagnostic rejection example.

These cases demonstrate why retrieval, reranking, and bridge selection should not be treated as a single mapping step. In a retrieval failure such as `TH_001_1`, no downstream evidence function can recover the reviewed target because it is absent from the candidate pool. In a selection failure such as `TH_035_1`, the acceptable target is available but another candidate receives the final selection. In `TH_002_1`, retrieval also succeeds, but the configuration abstains because the final operating rule does not materialise a bridge. These cases require different improvements: stronger candidate generation for retrieval failures, better ranking or evidence integration for selection failures, and better calibrated selection criteria for abstention behaviour.

5.4. Transferability and Answers to the Research Questions

The Spanish run demonstrates procedural reuse rather than quantitative cross-language equivalence. The same target ontology, multilingual representation, candidate-union procedure, evidence schema, bridge representation, and provenance structure were reused after replacing the source-language resources. The run processed 250 source nodes, including 200 test nodes, and produced 200 final prediction records. Because the prediction identifiers could not be defensibly linked to the separate 64-node pilot reference resource, no Spanish precision, recall, F1, or cross-language accuracy values are reported; reconstructing evaluation labels after observing the predictions would risk post-hoc bias.

For RQ1, the configurations exhibited different coverage–selectivity trade-offs at their validation-selected operating points. Lexical achieved the highest F1, Combined the highest precision, and Complete emitted fewer bridges without improving F1. These results characterise end-to-end configuration behaviour rather than a matched-coverage estimate of the isolated contribution of individual evidence signals. Candidate-union Recall@10 of 0.742 further identifies retrieval as an upstream constraint. For RQ2, 17/36 Complete bridges were exact matches, whereas 32/36 were connected within six hierarchy edges; 15/19 incorrect bridges also met this criterion, including 13 within one or two hops. Structural proximity is therefore a complementary graph property rather than evidence of semantic equivalence. For RQ3, the construction procedure could be executed after replacing the Thai resources with the controlled Spanish source resource, supporting procedural transferability but not equivalent cross-language mapping effectiveness.

5.5. Limitations and Future Directions

The evaluation is limited by the small Thai graph, the author-verified silver reference set, and the 26-node validation split, which may make the selected operating points sensitive to individual cases. Candidate retrieval is also a bottleneck: 25 of the 97 mapped test targets were absent from the union of the two Top-10 candidate lists and could not be recovered by later stages. In addition, only one intentional no-match case was available, so the observed rejection should not be interpreted as general open-world abstention performance. Larger independently reviewed datasets with more diverse unmapped cases are therefore required.

The reported configuration comparison also combines ranking quality with different levels of abstention: Lexical emitted 98 bridges whereas Complete emitted 36. Future evaluation should therefore include threshold-sweep or matched-coverage analyses. Structural evidence is likewise preliminary: local adjacency is relation-agnostic, target proximity uses an undirected `is-a` hierarchy with a predefined six-hop bound, and many incorrect bridges were still structurally close. The present audit uses incorrect emitted bridges as an internal comparator but does not include a random-target distance baseline. Relation-aware evidence and stronger structural comparators should be investigated.

The Spanish experiment is further limited by identifier provenance and its controlled construction from Spanish SNOMED CT descriptions. Future work should use an independently developed source graph with identifier-consistent reviewed mappings. Evaluation should also extend beyond SNOMED CT to additional terminologies and ontology releases, and examine whether retained bridge evidence improves downstream querying, interoperability, curator review, and human-audit workflows.

6. Conclusion

This paper presented a multilingual workflow for constructing and materialising evidence-bearing semantic bridge relations between local clinical knowledge graphs and standard clinical ontologies. The workflow separates candidate retrieval, evidence-based reranking, thresholded selection or abstention, and persistent graph materialisation. A materialised SEMANTIC_BRIDGE records an automatically selected correspondence together with its reranking score, component evidence, provenance, source language, target ontology, and ontology release; it does not by itself assert semantic equivalence.

In the Thai–SNOMED CT evaluation, candidate-union Recall@10 reached 0.742. The Lexical configuration achieved the highest F1 (0.338), while Combined achieved the highest precision (0.541). Complete emitted 36 bridges with precision 0.472, recall 0.175, F1 0.256, and coverage 0.367. The ontology audit showed that 32/36 Complete bridges were connected to an acceptable reviewed target within six undirected active inferred `is-a` edges, but only 17 were exact matches. Moreover, 15/19 incorrect bridges also satisfied the connectivity criterion. Structural proximity should therefore be treated as a complementary graph property rather than as evidence of semantic equivalence. The Spanish experiment further showed that the language-independent construction procedure can be reused after replacing source-language resources, although the current identifier provenance does not support a quantitative cross-language accuracy comparison.

The contribution is therefore not a replacement mapping standard, but an explicit workflow for deciding when a retrieved correspondence should be materialised as a persistent and inspectable graph relation. By retaining the evidence and provenance associated with that decision, semantic bridge relations provide a basis for more auditable cross-resource clinical knowledge integration. Future work should evaluate larger independently reviewed datasets, stronger multilingual retrieval and relation-aware structural evidence, additional target ontologies, and downstream graph querying and human-audit tasks.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language editing, structural revision, terminology-review support, and code/audit support, and QuillBot for grammar and spelling checks. All experimental decisions, analyses, reported results, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the publication.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37. doi:10.1145/3447772.

[2] O. Bodenreider, The unified medical language system (UMLS): Integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[3] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of Biomedical Informatics 137 (2023) 104252. doi:10.1016/j.jbi.2022.104252.

[4] K. S. Kalyan, S. Sangeetha, Medical concept normalization in user-generated texts by learning target concept embeddings, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Online, 2020, pp. 18–23. doi:10.18653/v1/2020.louhi-1.3.

[5] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, E. Fosler-Lussier, Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets, Journal of the American Medical Informatics Association 28 (2021) 516–532. doi:10.1093/jamia/ocaa269.

[6] B. Portelli, S. Scaboro, G. Serra, Improving multi-lingual medical term normalization to address the long-tail problem, in: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Hong Kong, China, 2023, pp. 809–818. URL: https://aclanthology.org/2023.paclic-1.81/.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 878–891. URL: https://aclanthology.org/2022.acl-long.62/. doi:10.18653/v1/2022.acl-long.62.

[8] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, World Wide Web Consortium, 2009. URL: https://www.w3.org/TR/skos-reference/.

[9] N. Matentzoglu, J. P. Balhoff, S. M. Bello, C. Bizon, M. Brush, T. J. Callahan, C. G. Chute, W. D. Duncan, C. T. Evelo, D. Gabriel, J. Graybeal, A. Gray, B. M. Gyori, M. Haendel, H. Harmse, N. L. Harris, I. Harrow, H. B. Hegde, A. L. Hoyt, C. T. Hoyt, D. Jiao, E. Jiménez-Ruiz, S. Jupp, H. Kim, S. Koehler, T. Liener, Q. Long, J. Malone, J. A. McLaughlin, J. A. McMurry, S. Moxon, M. C. Munoz-Torres, D. Osumi-Sutherland, J. A. Overton, B. Peters, T. Putman, N. Queralt-Rosinach, K. Shefchek, H. Solbrig, A. Thessen, T. Tudorache, N. Vasilevsky, A. H. Wagner, C. J. Mungall, A simple standard for sharing ontological mappings (SSSOM), Database 2022 (2022) baac035. URL: https://doi.org/10.1093/database/baac035. doi:10.1093/database/baac035.

[10] T. Lebo, S. Sahoo, D. McGuinness, PROV-O: The PROV Ontology, W3C Recommendation, World Wide Web Consortium, 2013. URL: https://www.w3.org/TR/prov-o/.

[11] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for eHealth, in: Studies in Health Technology and Informatics, volume 121, IOS Press, 2006, pp. 279–290.

[12] S. Hertling, H. Paulheim, OLaLa results for OAEI 2023, in: Proceedings of the 18th International Workshop on Ontology Matching, volume 3591 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 170–177. URL: https://ceur-ws.org/Vol-3591/oaei23_paper7.pdf.

[13] H. Babaei Giglou, J. D'Souza, S. Auer, M. Sanaei, OntoAligner meets knowledge graph embedding aligners, in: Proceedings of the 20th International Workshop on Ontology Matching, volume 4144 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 39–51. URL: https://ceur-ws.org/Vol-4144/om2025-LTpaper4.pdf.

[14] SNOMED International, SNOMED CT mapping reference set, SNOMED CT Documentation, n.d. URL: https://docs.snomed.org/snomed-international-documents/snomed-ct-glossary/m/mapping-reference-set, accessed: 24 August 2026.

[15] C. T. Hoyt, K. Karis, B. M. Gyori, Assembly and reasoning over semantic mappings at scale for biomedical data integration, Bioinformatics 41 (2025) btaf542. URL: https://doi.org/10.1093/bioinformatics/btaf542. doi:10.1093/bioinformatics/btaf542.

[16] M. Schaeffer, M. Sesboüé, L. Charbonnier, N. Delestre, J.-P. Kotowicz, C. Zanni-Merk, On the pertinence of LLMs for ontology learning, in: Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation, volume 3874 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1–18. URL: https://ceur-ws.org/Vol-3874/paper1.pdf.

[17] V. K. Kommineni, B. König-Ries, S. Samuel, Towards the automation of knowledge graph construction using large language models, in: Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation, volume 3874 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 19–34. URL: https://ceur-ws.org/Vol-3874/paper2.pdf.

[18] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[19] SNOMED International, SNOMED CT release file specification, SNOMED CT Document Library, 2025. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-release-file-specification, accessed: 20 July 2026.