Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Granularity Blind Spot in the Evaluation of SNOMED CT Concept Normalization

Yejin Jo
Regional Emergency Medical Center, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine

Yoojin Nam
Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center

YangWon Kim
Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine

Eunji Kim
Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine

Pa Hong
papa.hong@samsung.com

Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine

Research Article

Keywords: concept normalization, medical entity linking, semantic annotation, is-a hierarchy, subsumption, semantic similarity, SNOMED CT, biomedical terminology, evaluation methodology

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10602575/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Granularity Blind Spot in the Evaluation of SNOMED CT Concept Normalization

Yejin Jo[1], Yoojin Nam[2,3], YangWon Kim[4], Eunji Kim[4], Pa Hong[4,5,*]

[1] Regional Emergency Medical Center, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, 158 Paryong-ro, Masanhoewon-gu, Changwon-si, Gyeongsangnam-do 51353, Republic of Korea [2] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea [3] Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea [4] Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, 158 Paryong-ro, Masanhoewon-gu, Changwon-si, Gyeongsangnam-do 51353, Republic of Korea [5] Department of Digital Health, Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea

* Corresponding author: Pa Hong (papa.hong@samsung.com)

ORCID — Yejin Jo 0009-0002-0158-7950 · Yoojin Nam 0000-0001-8565-1360 · YangWon Kim 0009-0004-1123-9858 · Eunji Kim 0009-0003-3075-9683 · Pa Hong 0000-0001-5495-5230

Author emails — Yejin Jo (cong115@naver.com) · Yoojin Nam (eugene8998@gmail.com) · YangWon Kim (yw1117.kim@samsung.com) · Eunji Kim (ejk6111@gmail.com) · Pa Hong (papa.hong@samsung.com)

Abstract

Background. Clinical concept normalization maps free-text mentions to terminology codes but is usually evaluated with aggregate exact match and a binary seen/novel split. These summaries do not distinguish mentions recorded as finer-grained than their assigned code. Using native granularity labels in three Spanish SNOMED CT corpora, we tested whether exact-match error incidence differs between NARROW and EXACT reference strata across up to five lexical, neural and generative systems.

Results. Systems failed to reproduce the corpus-assigned code under exact match more often for NARROW mentions (finer-grained than their assigned code) than for EXACT mentions (panel-weighted Mantel–Haenszel risk ratio 1.72, 95% CI 1.63–1.83). The Wu–Palmer-minus-exact credit gap was larger on NARROW mentions in all 14 system×benchmark cells (geometric mean ratio 1.87, 95% CI 1.71–2.05), driven by error incidence rather than a consistent difference in conditional error proximity. In the pooled Qwen analysis, sibling and cousin predictions contributed 56–67% of summed Wu–Palmer credit among non-exact predictions. Under a

stricter same-lineage criterion, the uplift beyond exact match was larger for NARROW mentions in 9 of 14 cells. The larger graded gap retained its direction under alternative metrics, candidate spaces and xMEN. Graph-refined novelty classified 825 of 982 code-novel DisTEMIST mentions (84.0%) as ancestor-seen; the remaining 157 were fully isolated (6.3% of all 2,507 mentions). Majority-vote consensus in the blinded three-rater audit agreed with the corpus labels for 130 of 150 mentions ($\kappa = 0.73$).

Conclusions. Aggregate exact match obscures a benchmark-relative difference between reference-granularity strata. For benchmarks with granularity labels, granularity-stratified exact-match error rates should be reported first; hierarchy-aware scores should be paired with relation types.

Keywords: concept normalization; medical entity linking; semantic annotation; is-a hierarchy; subsumption; semantic similarity; SNOMED CT; biomedical terminology; evaluation methodology

Background

Clinical concept normalization, or medical entity linking, maps free-text mentions to concepts in a terminology such as the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT). By grounding free text in a formal terminology, normalization outputs can affect phenotyping, cohort selection, coding, decision support and cross-institutional data exchange. SNOMED CT is a multilingual reference terminology whose concepts form a poly-hierarchical is-a graph rather than an unordered code list.

Exact match answers whether the predicted and gold codes are identical, but two non-exact predictions can occupy very different positions in the SNOMED CT hierarchy. Hierarchy-aware metrics locate those errors, yet aggregate exact and hierarchy-aware scores still conceal whether performance differs with the specificity of the reference annotation. Both aggregate scoring and the binary seen/novel split can collapse relevant hierarchy structure: reference granularity in the first case and training-to-test graph connectivity in the second. Reference-granularity stratification is the primary analysis; graph-refined novelty is a secondary reporting extension.

Native corpus annotators flag some mentions as NARROW: more specific than the assigned code. In an observed SympTEMIST case, “bladder wall thickening” was assigned the broader concept *Disorder of urinary bladder*. The evaluated LLM reranker returned the direct parent of the assigned concept, *Urinary bladder finding*: the prediction scored zero under exact match and remained close in the hierarchy, although it moved further from the mention’s specificity. We call the within-stratum difference between mean Wu–Palmer and mean exact match the credit gap. We use this gap as a diagnostic by separating error incidence from conditional error proximity, then report relation types because recovered credit can be lateral rather than on the same is-a lineage.

Related work

Hierarchy-aware evaluation of biomedical entity linking rests on an established metric literature: least-common-ancestor similarity [1], graph distance [2], and count-preserving hierarchical evaluation for SNOMED-style multi-label coding [3]. Biomedical linkers such as SapBERT [4] and BioSyn [5], including cross-lingual toolkits such as xMEN [6], are commonly evaluated with aggregate accuracy and seen/novel status. Standard in-batch InfoNCE [7] provides the contrastive objective used in our fine-tuning probe.

Most recently, Xu, Nenadic and Stevens [8] proposed a specificity-sensitive framework that classifies predictions by their logical relation to a fixed gold concept. We use their relation typology but examine the opposite side of the annotation: whether the source corpus records the mention itself as finer-grained than its assigned code. Prediction relation and reference granularity are distinct axes: one locates an error, whereas the other identifies a reference stratum in which exact-code reproduction may differ. We test this reference-side axis across three Spanish SNOMED CT benchmarks and multiple systems.

Contribution

Our central empirical finding is that mentions recorded as finer-grained than their assigned code have higher exact-match error rates across the sampled systems, entity types and benchmarks, a difference hidden by aggregate reporting. We separate error incidence from conditional error proximity, identify which ontology relations carry hierarchy-aware credit and test the pattern across evaluation choices. As a secondary analysis, we divide novel mentions into ancestor-seen and fully isolated groups. The released scorer, stratifications and identifier-free per-item metrics support these analyses (Figure 1).

Methods

Data and benchmarks

We drew up to 800 eligible rows per Spanish SNOMED CT benchmark once with a fixed seed, preserving each benchmark's native EXACT/NARROW prevalence. De-duplication to unique (mention, gold) pairs left 678 DisTEMIST, 625 MedProcNER and 800 SympTEMIST rows; these fixed samples were held constant across systems and metrics. The common 800-row target set a uniform LLM evaluation budget. Eligibility was restricted to EXACT and NARROW rows before sampling; stratum sizes are summarized in Table 4. The released identifier-free rows, code and seeds support result recalculation.

Two secondary analyses used different samples. We evaluated an 800-mention English MedMentions sample by depth band and used the full DisTEMIST test set (n = 2,507) for the overlap analysis.

- **DisTEMIST** disease mentions [9]: system-training/candidate pool 4,935 single-code rows / 2,202 unique codes; test 2,507 eligible EXACT/NARROW rows / 1,391 unique codes; sample n = 678 (EXACT 443 / NARROW 235).
- **MedProcNER/ProcTEMIST** procedure mentions [10]: sample n = 625 (EXACT 491 / NARROW 134); candidate space 1,687 concepts.
- **SympTEMIST** symptom/finding mentions [11]: sample n = 800 (EXACT 603 / NARROW 197); candidate space 1,810 concepts.
- **MedMentions** English mixed-type mentions [12], mapped through UMLS [13]: sample n = 800. Because it lacks EXACT/NARROW labels, we use depth bands (shallow ≤ 2 / mid = 3 / deep ≥ 4) as a depth-proxy cross-check.

Agreement was quantified with Cohen's κ [14] and interpreted using the Landis–Koch scale [15]. SNOMED CT provides the reference terminology [16] and is-a hierarchy semantics [17].

Granularity labels

The Spanish corpora provide human granularity labels in the semantic_rel field: EXACT denotes matched specificity; NARROW denotes a mention more specific than the assigned code. We refer to the code assigned in each corpus as the gold code. DisTEMIST used two clinical specialists and physician review (reported pairwise agreement for disease mentions, 82.3) [18]; MedProcNER used two clinical experts and revision by a third physician (reported total pairwise agreement, 81.2) [19]. These values were not specific to EXACT/NARROW, and SympTEMIST did not report corresponding annotation details, so we conducted a separate granularity-label reproducibility audit.

Rows outside EXACT/NARROW were excluded from stratified evaluation. Single-code rows from other strata could still contribute to system training or candidate-space construction; COMPOSITE, NO_CODE and concatenated multi-code rows were not evaluated.

Terminology and language

The released granularity labels were analyzed as provided and reflect the corpora's annotation-time SNOMED CT editions (DisTEMIST 2022; MedProcNER and SympTEMIST 2023). All hierarchy-based scoring used terminology labels and is-a ancestors retrieved through OLS4 on 18–19 July 2026. Because the API did not expose an RF2 release identifier, we fixed the derived maps by retrieval date and SHA-256 digest. We defined depth as the longest root-to-node path and the least common ancestor (LCA) as the deepest common ancestor.

Spanish mentions were translated with Qwen2.5-32B-Instruct before matching English labels; MedMentions required no translation. Additional file 1 gives the decoding and depth-sensitivity details.

Candidate spaces and ontology release

The primary corpus-relative inventory contains the training gold codes and the gold codes of all evaluated test items. For MedProcNER and SympTEMIST, a complete benchmark-wide test candidate catalogue was unavailable; only the evaluated test golds were added. We reran all systems in a harder training-only inventory after removing test-only codes. Candidate counts, frozen-map construction and unresolved-code rules are described in Additional file 1. Table 1 reports the primary corpus-relative arm.

Normalization systems

We scored five lexical, neural and generative systems:

- **Lexical (TF–IDF):** character n-gram (2–4) cosine similarity.
- **SapBERT:** cambridgeltl/SapBERT-from-PubMedBERT-fulltext classification-token (CLS) embedding cosine retrieval [4].
- **Fine-tuned SapBERT:** standard in-batch InfoNCE [7] on DisTEMIST and SympTEMIST. Hyperparameters are in Additional file 1.
- **LLM reranker:** Qwen2.5-32B-Instruct (Apache-2.0; revision 5ede1c97) greedily returned an ordered list of five identifiers from the top-20 lexical candidates. The first valid identifier supplied the top-1 scores and the list supplied Recall@5; if none was valid, lexical order was used. Its ranking was bounded by lexical Recall@20.
- **LLM generator:** the same model independently proposed a concept name without a candidate list; the proposal was fuzzy-matched (Weighted Ratio) to the concept space.

LLM protocol. Both used fixed prompts and greedy decoding on one 80-GB graphics processing unit (GPU). Prompt recipes, code, seeds, model revisions, fallback counts and environment details are provided in Additional file 1 and the release.

Use of large language models in manuscript preparation. A large language model assisted with language editing and with code writing for the rescoring and statistical analyses. The authors checked and edited all such output and are responsible for the content of the paper. No AI tool is listed as an author. This is separate from the large language model evaluated as a normalization system in this study (Qwen2.5-32B-Instruct), described above.

Metrics

We report exact match, is-a consistency, Recall@5 and Wu–Palmer = 2·depth(LCA) / (depth(pred) + depth(gold)) [1]. Is-a consistency credits a prediction only when it is the gold code, an ancestor of it, or a descendant of it. It therefore separates same-lineage near-misses from lateral (sibling) errors; off-map pairs score zero. We also tested normalized inverse LCA path distance and ancestor-set Jaccard, inspired by but distinct from CoPHE [3]. Root handling and metric definitions are in Additional file 1.

Credit gap. For a system and stratum, the credit gap is mean(Wu–Palmer) minus mean(exact match). It describes hierarchy proximity, not clinical correctness. Throughout, "error" means failure to reproduce the corpus-assigned code under exact match. The NARROW/EXACT ratio is gap(NARROW) / gap(EXACT); a ratio > 1 marks a larger score difference on NARROW mentions.

Overlap (seen/novel) stratification

Each gold code is code-seen, ancestor-seen, or fully-isolated relative to benchmark training codes. Ancestor-seen means that the gold code is unseen but at least one proper ancestor is training-seen; fully isolated means that neither the gold code nor any proper ancestor is training-seen. The last two buckets form the conventional "novel" set.

Fine-tuning-exposure buckets

Fine-tuning gains were grouped by code exposure in the fine-tuning pairs. These buckets differ from the training-overlap split above; details are in Additional file 1.

Robustness and validity analyses

We tested metric and depth choices, an NLLB-200 translation check, training-only candidate spaces, xMEN [6], and the public ieeta-pt MedProcNER runs. We also decomposed all 14 cells into error incidence and conditional proximity and classified errors by ontology relation.

For the human reproducibility audit, three clinical annotators independently re-annotated a balanced 150-mention sample (50 per benchmark; 25 corpus-EXACT and 25 corpus-NARROW) while blinded to native labels and system outputs. We report per-annotator Cohen's κ, inter-annotator Fleiss κ and majority-vote consensus κ against the native labels. The balanced sample was designed to estimate agreement rather than population label frequencies. Additional file 1 gives the full protocols and outputs.

Statistical analysis

Inference used archived per-mention outputs. Each of the 14 system×benchmark combinations formed a stratum, with NARROW versus EXACT as the exposure and failure to reproduce the assigned code as the outcome. The

primary summary is the panel-weighted Mantel–Haenszel error-incidence risk ratio; we also report the corresponding odds ratio for comparison with earlier work. Their 95% intervals use a fixed-seed bootstrap that resamples mentions within each benchmark and applies the same indices to every system (B = 10,000; seed 12345). The multiplicative decomposition instead uses the geometric mean of the 14 per-cell incidence ratios because the identity holds between geometric means, not pooled estimators. We used exact McNemar tests for fine-tuning comparisons and Cohen's κ and Fleiss κ for agreement.

We assessed heterogeneity with exploratory Breslow–Day tests, holding the system fixed and using disjoint benchmarks as strata, and applied a Holm family-wise sensitivity across the five tests. The NARROW/EXACT gap ratio received a fixed-seed stratum-wise bootstrap interval (B = 2,000; seed 20260805); per-cell intervals resampled the two strata independently, whereas the geometric-mean interval shared indices across systems within each benchmark. We also repeated the incidence analysis among mentions whose gold concept had at least one descendant in the frozen candidate-space-plus-ancestor graph. Code and outputs are deposited (see Availability of data and materials).

To test whether the granularity label acts as a proxy for lexical or ontological difficulty, we fitted a modified-Poisson model of exact-match error adjusted for mention length, gold-code training frequency, local candidate-space density, gold-code depth and system, with cluster-robust variance on source document. The model was confined to SympTEMIST because it was the only benchmark whose fixed analytic rows retained source-row alignment.

Results

Overall performance across four benchmarks

Table 1 reports each benchmark's overall scores. Off-the-shelf SapBERT and the LLM reranker were closely matched on DisTEMIST (Wu–Palmer 0.826 vs 0.823; exact 0.637 vs 0.633), with wider separation on MedProcNER and SympTEMIST.

NARROW mentions have higher exact-match error incidence

Within each stratum, the credit gap factors into error incidence and conditional error proximity:

$$G_s = P_s(\text{error}) \times \text{Wu}\bar{\text{P}}_{\text{error},s},\ s \in \{\text{EXACT, NARROW}\}$$

where $\text{Wu}\bar{\text{P}}_{\text{error},s}$ is the mean Wu–Palmer similarity among non-exact predictions in stratum s. The NARROW/EXACT gap ratio is therefore the product of the corresponding incidence and conditional-proximity ratios; a component ratio above 1 indicates a larger NARROW component.

The primary panel-weighted Mantel–Haenszel error-incidence risk ratio was 1.72 (95% CI 1.63–1.83); the corresponding secondary odds ratio was 3.89 (95% CI 3.33–4.56; Figure 2b). The incidence ratio exceeded 1 in all 14 system×benchmark cells. For the multiplicative decomposition, the equal-cell geometric mean incidence ratio was 1.81 (95% CI 1.70–1.94), while the conditional-proximity ratio exceeded 1 in 6 cells and had a geometric mean of 1.03; unrounded values yielded the 1.87× total ratio (1.81 × 1.03; Figure 2a). The equal-cell mean conditional-proximity difference was +0.007 (95% CI −0.022 to +0.037) and changed sign across benchmarks (Additional file 1). Thus, failure to reproduce the assigned code was more frequent in NARROW mentions, whereas conditional error proximity showed no consistent cross-benchmark direction.

In the SympTEMIST analysis with source-linked covariates, the granularity risk ratio was 1.77 (95% CI 1.61–1.94) before adjustment and 1.55 (95% CI 1.40–1.72) after adjustment for mention length, gold-code training frequency, candidate-space sibling density, gold-code depth and system (4,000 mention×system rows; 213 document clusters).

All 14 cell-specific odds ratios, reported as secondary effect measures, exceeded 2 (range 2.20–15.45), but the magnitude varied. In exploratory Breslow–Day tests, only the LLM generator showed heterogeneity after Holm correction across five systems (adjusted $P = 0.001$; Additional file 1).

What hierarchy-aware scoring adds: the granularity credit gap

The NARROW credit gap exceeded the EXACT gap in all 14 cells (Figure 3; Table 2). Additive-gap bootstrap intervals excluded zero in 13 of 14 cells, and the geometric-mean NARROW/EXACT ratio was 1.87 (95% CI 1.71–2.05). The sole exception by either interval criterion was the MedProcNER LLM generator (ratio 1.10, 95% CI 0.95–1.26).

What the errors are: relation-type composition

We classified each non-exact prediction as ancestor (under-specific), descendant (over-specific), sibling/cousin (lateral; termed *orthogonal* in [8]), or unresolved. For a descriptive Qwen-family profile, we pooled the reranker and generator (2,397 errors); per-system shares are reported separately.

Sibling/cousin errors accounted for 59–87% of error counts across the reranker and generator cells; descendants accounted for 2–23% and ancestors for 5–21% (Figure 4). In the pooled Qwen profile, descendant shares of error counts differed by only +2.4, −1.2 and +0.3 percentage points (NARROW − EXACT) across DisTEMIST, MedProcNER and SympTEMIST. The larger NARROW gap therefore did not arise mainly from over-specific predictions.

In the pooled Qwen profile, sibling/cousin errors contributed 58%, 56% and 67% of summed Wu–Palmer credit among non-exact predictions. In a stricter binary is-a sensitivity that excludes lateral credit, we defined uplift as mean is-a consistency minus mean exact match within each stratum. This uplift was larger for NARROW mentions in 9 of 14 cells, and the bootstrap 95% CI for the NARROW-minus-EXACT difference excluded zero in 6 of 14. The universal graded gap therefore did not extend to ancestor/descendant consistency (Additional file 1).

The NARROW label: a three-annotator human reproducibility check

Three clinical annotators, blinded to native labels and system outputs, independently re-annotated the balanced 150-mention sample. Inter-annotator agreement was moderate (Fleiss $\kappa = 0.519$, 95% CI 0.41–0.62), and majority-vote consensus agreed with the corpus label on 130 mentions (86.7%; $\kappa = 0.73$, 95% CI 0.61–0.84; Table 3).

Robustness

Model adaptation and metrics. Fine-tuning improved every metric (DisTEMIST exact 0.637→0.677; Wu–Palmer 0.826→0.857) while leaving the credit gap intact. The ratio increased from 2.36× to 2.66× on DisTEMIST and decreased from 3.63× to 3.44× on SympTEMIST. The ancestor-seen gain among codes unseen during fine-tuning (+6.1 points; n = 261) did not differ significantly from the code-seen gain (Additional file 1). Normalized inverse LCA path distance and ancestor-set Jaccard retained NARROW > EXACT in all 14 cells. On DisTEMIST, system ordering was identical across exact match, Wu–Palmer, inverse LCA path distance and ancestor-set Jaccard; both tested node-depth conventions likewise retained ratios above one (Additional file 1).

Depth proxy and candidate space. In MedMentions, credit gaps increased from shallow to deep bands for every system. Because MedMentions lacks granularity labels and depth is partly coupled to Wu–Palmer, this is a cross-check rather than a direct replication (Additional file 1). Removing every test-only code reduced absolute accuracy and attenuated the pattern, but ratios remained ≥ 1 in all 14 cells. Additive-gap intervals excluded zero in 10 of 14 cells, versus 13 under the primary corpus-relative arm (Additional file 1).

Frozen-map leaf restriction. Gold concepts with at least one descendant in the candidate-space-plus-ancestor graph occurred in 46.8% of NARROW versus 44.2% of EXACT mentions on DisTEMIST, 48.5% versus 48.3% on MedProcNER, and 43.7% versus 30.3% on SympTEMIST. After restriction to these concepts, the error-incidence odds ratio was 6.01 (95% CI 4.79–7.69), compared with 3.89 (95% CI 3.33–4.56) overall. The contrast therefore persisted after leaves in the frozen map were excluded (Additional file 1).

External implementations and graph coverage. On separate stratified 800-mention draws, xMEN [6] showed the same direction on all three Spanish benchmarks (ratios 3.72, 2.05 and 2.49). Its candidate-generation mode and corpus-relative spaces differed from the primary analysis, so we compare the pattern rather than absolute scores (Figure 5). Because ieeta-pt used the full gazetteer, only 56% of EXACT and 17% of NARROW predictions resolved in our graph (Additional file 1). On the resolvable domain its runs showed NARROW/EXACT ratios of 2.84–3.20, whereas the ratio fell below 1 over the full matched output because unresolved NARROW predictions were zero-scored (Figure 5). This contrast shows how differential graph coverage changes the estimated ratio.

The flat seen/novel split hides hierarchical structure

As a secondary reporting extension, we refined conventional code-level novelty by whether an unseen gold code had a training-seen ancestor. In the full DisTEMIST test set (n = 2,507), 60.8% of mentions were code-seen and 39.2% novel. Of the latter, 84.0% had a seen ancestor (32.9% of all mentions); only 6.3% (n = 157) were fully isolated (Additional file 1). Distance-band results are shown in Additional file 1.

Discussion

Across the sampled Spanish benchmarks, NARROW mentions had higher benchmark-relative exact-match error incidence (panel-weighted risk ratio 1.72, 95% CI 1.63–1.83), while the conditional-proximity difference was small overall and changed direction across benchmarks. The main pattern therefore concerns whether systems reproduce the gold code, not where their non-exact predictions consistently land; its direction held in every cell, with varying magnitude. These estimates describe benchmark-target reproduction, not clinical failure.

Within that benchmark-relative interpretation, these results extend two established lines of evaluation. LCA similarity, graph distance and CoPHE locate or count hierarchy-aware disagreement [1–3], while Xu et al. classify each prediction's logical relation to a fixed gold code [8]. Neither approach stratifies performance by a property of the reference annotation. We add reference-granularity stratification, separate error incidence from conditional error proximity, and use relation types to distinguish in-lineage from lateral credit.

The blinded three-rater audit quantified the reproducibility of the reference-granularity axis. Majority-vote consensus agreed with the native corpus label for 86.7% of mentions ($\kappa = 0.73$), with moderate inter-annotator agreement at the EXACT/NARROW boundary.

Granularity-stratified exact-match rates and hierarchy-aware scores answer different questions: the former should be reported first, and the latter added with relation types when semantic location matters. Applying this approach elsewhere will require comparable granularity annotation.

Limitations

The three Spanish corpora were produced within the same shared-task series using related annotation conventions, so an independently annotated non-Spanish benchmark remains an important test of generalizability. In SympTEMIST, attenuation after covariate adjustment indicates that measured difficulty explains part, but not all, of the association.

Hierarchy proximity does not establish clinical acceptability, and some NARROW mentions may require post-coordination rather than a finer pre-coordinated concept [20]. The mention bootstrap preserved repeated scoring across systems but did not model clustering within documents, so its intervals may be anti-conservative.

Conclusions

In the sampled Spanish benchmarks, systems failed to reproduce the corpus-assigned code more often for NARROW than EXACT mentions, whereas conditional error proximity varied in direction by benchmark. An aggregate exact-match score therefore obscures a reference-side difference that granularity-stratified rates expose. For benchmarks with granularity labels, granularity-stratified exact-match rates should be reported first and, when needed, hierarchy-aware scores paired with relation types. Separately, seen/novel reporting can distinguish code-seen, ancestor-seen and fully isolated mentions.

List of abbreviations

CI: confidence interval; CLEF: Conference and Labs of the Evaluation Forum; CoPHE: count-preserving hierarchical evaluation; InfoNCE: information noise-contrastive estimation; LCA: least common ancestor; LLM: large language model; NLLB: No Language Left Behind; OLS4: Ontology Lookup Service, version 4; OR: odds ratio; RR: risk ratio; SapBERT: self-alignment pretraining for biomedical entity representations; SNOMED CT: Systematized Nomenclature of Medicine, Clinical Terms; TF–IDF: term frequency–inverse document frequency; UMLS: Unified Medical Language System; xMEN: cross-lingual medical entity normalization (toolkit)

Declarations

Ethics approval and consent to participate

Not applicable. This study analyzed publicly released, de-identified shared-task corpora (DisTEMIST, MedProcNER/ProcTEMIST, SympTEMIST and MedMentions). The re-annotation collected task labels, not personal, health or identifying information. No patient-level data or animal subjects were collected by the authors.

Consent for publication

Not applicable.

Availability of data and materials

The four source corpora are publicly available: DisTEMIST (https://doi.org/10.5281/zenodo.6408476) [9], MedProcNER/ProcTEMIST (https://doi.org/10.5281/zenodo.7929830) [10], SympTEMIST (https://doi.org/10.5281/zenodo.8413866) [11] and MedMentions [12].

The identifier-free derived datasets, labels, code, aggregate outputs and figures are archived as version 1.0.8 in Zenodo (https://doi.org/10.5281/zenodo.21805578), with a public development repository at https://github.com/honeia85/snomed [21]. The archive contains the scorer, both candidate-space stratifications, per-item metrics for all 14 system×benchmark cells, statistical outputs, robustness analyses, de-identified three-rater labels and annotation materials. Canonical CPU scripts recalculate the core results from the identifier-free rows.

The SNOMED International licence precludes redistribution of the frozen ontology maps, SNOMED CT identifiers and identifier-linked descriptions. The release provides map-generation code, SHA-256 digests of the 18–19 July 2026 derived maps and identifier-free rows (schema in Additional file 1, Table S5). Authorized users can regenerate the maps and verify their digests.

The descendant-availability restriction and adjusted incidence model additionally require, respectively, a licensed is-a graph and corpus annotation offsets; their code and outputs are included. Recall@5 is provided as an aggregate because ranked candidate lists are not redistributed. The release supports analysis-level reproduction; model decoding may vary across software and hardware environments.

The software is platform independent and requires Python 3.10 or later with NumPy, SciPy, statsmodels and pandas. Code is released under the MIT licence; author-generated identifier-free data, labels, figures and documentation are released under CC BY 4.0. Authorized SNOMED CT access is needed only to regenerate or query ontology maps.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

YJ conceived the study, contributed to its design, performed blinded granularity re-annotation, and contributed to drafting the manuscript. YN performed blinded granularity re-annotation, contributed to interpretation of the clinical granularity annotations, and critically revised the manuscript. YWK assisted with editing and revision of the manuscript. EK performed blinded granularity re-annotation. PH contributed to the study design, performed the analysis and rescoring, implemented the open-weight reproduction harness, released the code and data, drafted and finalized the manuscript, and is the corresponding author. All authors read and approved the final manuscript.

Acknowledgements

We thank the organizers of the DisTEMIST, MedProcNER and SympTEMIST shared tasks (Barcelona Supercomputing Center) for the gold-standard corpora, and the maintainers of the EBI Ontology Lookup Service.

References

1. Wu Z, Palmer M. Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics; 1994 Jun 27-30; Las Cruces, New Mexico, USA. Morristown (NJ): Association for Computational Linguistics; 1994. p. 133-8. https://doi.org/10.3115/981732.981751
2. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998. p. 265-84. https://doi.org/10.7551/mitpress/7287.003.0018
3. Falis M, Dong H, Birch A, Alex B. CoPHE: a count-preserving hierarchical evaluation metric in large-scale multi-label text classification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 907-12. https://doi.org/10.18653/v1/2021.emnlp-main.69

4. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg (PA): Association for Computational Linguistics; 2021. p. 4228-38. https://doi.org/10.18653/v1/2021.naacl-main.334
5. Sung M, Jeon H, Lee J, Kang J. Biomedical entity representations with synonym marginalization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg (PA): Association for Computational Linguistics; 2020. p. 3641-50. https://doi.org/10.18653/v1/2020.acl-main.335
6. Borchert F, Llorca I, Roller R, Arnrich B, Schapranow MP. xMEN: a modular toolkit for cross-lingual medical entity normalization. JAMIA Open. 2025;8(1):ooae147. https://doi.org/10.1093/jamiaopen/ooae147
7. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748 [cs.LG]. 2018. https://doi.org/10.48550/arXiv.1807.03748
8. Xu F, Nenadic G, Stevens R. A practical and nuanced framework for entity linking evaluation. J Biomed Semantics. 2026;17(1):5. https://doi.org/10.1186/s13326-025-00339-0
9. Miranda-Escalada A, Farré E, Gasco L, Lima S, Krallinger M. DisTEMIST corpus: detection and normalization of disease mentions in Spanish clinical cases [dataset]. Zenodo; 2022. https://doi.org/10.5281/zenodo.6408476
10. Lima López S, Farré Maduell E, Gascó Sánchez L, Krallinger M. MedProcNER/ProcTEMIST corpus: gold standard annotations for clinical procedures information extraction [dataset]. Zenodo; 2023. https://doi.org/10.5281/zenodo.7929830
11. Lima López S, Gascó Sánchez L, Farré E, Vigil Gimenez L, Krallinger M. SympTEMIST corpus: gold standard annotations for clinical symptoms, signs and findings information extraction [dataset]. Zenodo; 2023. https://doi.org/10.5281/zenodo.8413866
12. Mohan S, Li D. MedMentions: a large biomedical corpus annotated with UMLS concepts. In: Automated Knowledge Base Construction (AKBC 2019); 2019 May 20-22; Amherst (MA), USA. AKBC; 2019. https://doi.org/10.24432/C5G59C
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-70. https://doi.org/10.1093/nar/gkh061

14. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46. https://doi.org/10.1177/001316446002000104
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74. https://doi.org/10.2307/2529310
16. Richesson RL, Andrews JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. J Am Med Inform Assoc. 2006;13(5):536-46. https://doi.org/10.1197/jamia.M2093
17. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279-90.
18. Miranda-Escalada A, Gascó L, Lima-López S, Farré-Maduell E, Estrada D, Nentidis A, et al. Overview of DisTEMIST at BioASQ: automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In: Faggioli G, Ferro N, Hanbury A, Potthast M, editors. Working notes of CLEF 2022 — Conference and Labs of the Evaluation Forum; 2022 Sep 5-8; Bologna, Italy. Aachen: CEUR-WS.org; 2022. p. 179-203. (CEUR workshop proceedings; vol. 3180). https://ceur-ws.org/Vol-3180/paper-11.pdf. Accessed 5 Aug 2026.
19. Lima-López S, Farré-Maduell E, Gascó L, Nentidis A, Krithara A, Katsimpras G, et al. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In: Aliannejadi M, Faggioli G, Ferro N, Vlachos M, editors. Working notes of CLEF 2023 — Conference and Labs of the Evaluation Forum; 2023 Sep 18-21; Thessaloniki, Greece. Aachen: CEUR-WS.org; 2023. p. 1-18. (CEUR workshop proceedings; vol. 3497). https://ceur-ws.org/Vol-3497/paper-002.pdf. Accessed 5 Aug 2026.
20. Osborne JD, Neu MB, Danila MI, Solorio T, Bethard SJ. CUILESS2016: a clinical corpus applying compositional normalization of text mentions. J Biomed Semantics. 2018;9(1):2. https://doi.org/10.1186/s13326-017-0173-6
21. Jo Y, Nam Y, Kim YW, Kim E, Hong P. A granularity blind spot in the evaluation of SNOMED CT concept normalization. Version 1.0.8 [software and dataset]. Zenodo. 2026. https://doi.org/10.5281/zenodo.21805578

Tables

Table 1 Overall performance on four SNOMED CT benchmarks

Benchmark	System	n	Exact	Is-a	Wu–Palmer	Recall@5
DisTEMIST	lexical	678	0.488	0.606	0.732	0.671
DisTEMIST	SapBERT	678	0.637	0.736	0.826	0.816
DisTEMIST	SapBERT-ft	678	**0.677**	**0.785**	**0.857**	**0.873**
DisTEMIST	LLM reranker	678	0.633	0.742	0.823	0.763
DisTEMIST	LLM generator	678	0.301	0.422	0.559	—
MedProcNER	lexical	625	0.443	0.603	0.761	0.691
MedProcNER	SapBERT	625	**0.642**	**0.747**	**0.847**	**0.845**
MedProcNER	LLM reranker	625	0.602	0.746	0.838	0.755
MedProcNER	LLM generator	625	0.250	0.429	0.602	—
SympTEMIST	lexical	800	0.409	0.502	0.611	0.550
SympTEMIST	SapBERT	800	0.584	0.639	0.741	0.748
SympTEMIST	SapBERT-ft	800	**0.659**	**0.706**	**0.786**	**0.878**
SympTEMIST	LLM reranker	800	0.551	0.621	0.700	0.634
SympTEMIST	LLM generator	800	0.254	0.341	0.462	—
MedMentions	lexical	800	0.635	0.708	0.772	0.756
MedMentions	SapBERT	800	**0.851**	**0.896**	**0.915**	**0.943**
MedMentions	LLM reranker	800	0.688	0.757	0.812	0.769

Best Exact, Is-a, Wu–Palmer and Recall@5 values per benchmark are in bold. MedProcNER includes four systems and MedMentions three; neither includes fine-tuned SapBERT, and MedMentions also lacks archived generator predictions. The three Spanish benchmarks form the primary analysis; MedMentions is an English depth-proxy cross-check, and xMEN is reported separately. Recall@5 is undefined for the generator because it does not return a ranked candidate list. Scores are corpus-relative. Because complete benchmark-wide test candidate catalogues were unavailable for MedProcNER and SympTEMIST, their candidate spaces contain the evaluated golds and fewer distractors than DisTEMIST; absolute scores should therefore be compared within, rather than across, benchmarks.

Table 2 Per-system credit gap (Wu–Palmer − exact match) by native-corpus granularity label, three Spanish benchmarks

Benchmark	System	gap_EXACT	gap_NARROW	Ratio (N/E) [95% CI]	ΔG [95% CI]
DisTEMIST	SapBERT-ft	0.114	0.303	2.66 [2.01, 3.60]	0.189 [+0.135, +0.243]
DisTEMIST	SapBERT	0.128	0.303	2.36 [1.84, 3.06]	0.175 [+0.122, +0.227]
DisTEMIST	LLM reranker	0.136	0.291	2.14 [1.68, 2.77]	0.155 [+0.102, +0.210]
DisTEMIST	lexical	0.224	0.280	1.25 [1.01, 1.55]	0.057 [+0.005, +0.109]
DisTEMIST	LLM generator	0.227	0.315	1.39 [1.18, 1.66]	0.088 [+0.041, +0.134]
MedProcNER	SapBERT	0.166	0.350	2.10 [1.68, 2.62]	0.184 [+0.125, +0.242]
MedProcNER	LLM reranker	0.209	0.335	1.60 [1.30, 1.98]	0.126 [+0.065, +0.189]
MedProcNER	lexical	0.297	0.394	1.33 [1.11, 1.57]	0.097 [+0.035, +0.160]
MedProcNER	LLM generator	0.345	0.379	1.10 [0.95, 1.26]	0.034 [−0.019, +0.087]
SympTEMIST	SapBERT-ft	0.080	0.274	3.44 [2.62, 4.58]	0.195 [+0.145, +0.245]
SympTEMIST	SapBERT	0.095	0.346	3.63 [2.87, 4.66]	0.250 [+0.201, +0.299]
SympTEMIST	LLM reranker	0.113	0.258	2.28 [1.76, 2.88]	0.145 [+0.099, +0.193]
SympTEMIST	lexical	0.181	0.269	1.49 [1.22, 1.79]	0.089 [+0.043, +0.136]
SympTEMIST	LLM generator	0.195	0.249	1.28 [1.09, 1.52]	0.055 [+0.016, +0.094]

Credit gap = mean(Wu–Palmer) − mean(exact) within stratum. Ratio = gap(NARROW)/gap(EXACT); > 1 means hierarchy-aware scoring adds proportionally more score beyond exact match on NARROW mentions. Its geometric mean across the 14 cells is 1.87 (95% CI 1.71–2.05). ΔG is the additive NARROW-minus-EXACT gap. Fixed-seed mention bootstraps used B = 10,000 for ΔG and B = 2,000 for ratios; the geometric-mean interval shares sampled indices across systems within each benchmark. The MedProcNER LLM generator is the

only cell whose ratio interval includes 1 and whose ΔG interval includes zero. Rows within each benchmark are ordered by exact-match accuracy.

Table 3 Human granularity-label reproducibility

Annotation pass	Sample	Cohen's κ (95% CI)	Raw agreement
Study rater R1	balanced stratified, n = 150	0.61 (0.48–0.73)	80.7%
Study rater R2	balanced stratified, n = 150	0.63 (0.50–0.74)	81.3%
Study rater R3	balanced stratified, n = 150	0.67 (0.54–0.79)	83.3%

Across the three blinded annotators, inter-annotator Fleiss κ = 0.519 (95% CI 0.41–0.62); majority-vote consensus vs native corpus labels κ = 0.73 (95% CI 0.61–0.84).

Table 4 Evaluation-sample sizes by granularity stratum

Benchmark	EXACT	NARROW	Total
DisTEMIST	443	235	678
MedProcNER	491	134	625
SympTEMIST	603	197	800

Figures

Figure 1 Study overview. In the observed SympTEMIST example, the mention "bladder wall thickening" is assigned the broader concept *Disorder of urinary bladder*, which omits the specific wall thickening and is therefore labelled NARROW. The LLM reranker predicts the assigned code's direct parent, *Urinary bladder finding*: exact match is zero and Wu–Palmer remains high, although the prediction moves further from the mention's specificity. The pipeline applies this comparison across three Spanish benchmarks with EXACT/NARROW labels and MedMentions as an English depth proxy. Predictions from up to five normalization systems are evaluated by exact match and hierarchy-aware metrics on the SNOMED CT is-a graph. The granularity credit gap is Wu–Palmer minus exact match within each stratum. We decompose the gap, examine relation types, and run the system, metric, candidate-space, graph-coverage and overlap checks shown.

Figure 2 The granularity difference is incidence-dominated, not proximity-dominated. **(a)** The geometric-mean NARROW/EXACT gap ratio (1.87×) factors, on a log axis anchored at 1.0, into an equal-cell error-incidence component (1.81×) and a conditional-proximity component (1.03×). **(b)** The secondary panel-weighted Mantel–Haenszel error-incidence odds ratio is 3.89 (95% CI 3.33–4.56), with every per-system estimate above 2.7 (2.77–5.54); the primary risk-ratio estimate is 1.72 (95% CI 1.63–1.83). Exploratory unadjusted heterogeneity tests were below 0.05 for two systems; after Holm correction across five tests, only the LLM generator remained below 0.05.

Figure 3 Granularity credit gaps across 14 system×benchmark cells. Each row is one cell; the light point is the credit gap (Wu–Palmer − exact match) on EXACT mentions and the dark point on NARROW mentions, joined by the additive gap ΔG = gap_NARROW − gap_EXACT. NARROW exceeds EXACT in every cell; the

additive-gap bootstrap 95% CIs exclude zero in 13/14 cells (the hollow point marks the single exception, the MedProcNER LLM generator, ΔG +0.034 [−0.019, +0.087]). Rows are ordered by exact-match accuracy within each benchmark. Values are in Table 2.

Figure 4 Relation-type composition of non-exact predictions (pooled LLM reranker + generator), EXACT vs NARROW strata. Sibling/cousin ("orthogonal") predictions dominate both strata on all three benchmarks; the over-specific (descendant) share does not rise consistently from EXACT to NARROW (pooled NARROW−EXACT difference +2.4 percentage points on DisTEMIST, −1.2 on MedProcNER and +0.3 on SympTEMIST).

Figure 5 Selected cross-system checks. Each plotted row shows the NARROW/EXACT credit-gap ratio (points coloured by benchmark; ratio = 1 marks no gap): the in-house SapBERT reference (2.10–3.63), in-domain fine-tuning (2.66–3.44), the external xMEN toolkit (2.05–3.72) and a CLEF-2023 participant system (ieeta-pt, five-run range 2.84–3.20 on its graph-resolvable output). Alternative graph metrics, NLLB-200 translation and a harder training-only candidate space are reported in Additional file 1. The first three rows assess robustness across implementations; the participant row diagnoses sensitivity to graph coverage.

Additional files

Additional file 1 (.docx). Supplementary Methods, Supplementary Results, Tables S1–S5 and Figures S1–S6. Includes extended terminology, translation and candidate-space methods; system and fine-tuning details; metric implementation; robustness and inference protocols; re-annotation results; depth-proxy validity; fine-tuning exposure buckets; incidence/proximity decomposition; binary same-lineage sensitivity; descendant-availability restriction; exploratory Breslow–Day tests; the identifier-free per-mention schema; gap-ratio intervals; metric, depth and training-only sensitivity analyses; xMEN and participant audits; and extended limitations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx