Ontological Validation of Biomedical Topic Models: SNOMED CT Hierarchy Distance as an Automated Evaluation Metric

Ilan S. Rubinfeld[1] **Sami Zaidi**[1] **Milosh Djuric**[1] **Mouhammad Halabi**[1]
Loay Kabbani[1] **Alex Shepard**[1] **Mohammad Ghassemi**[2]
[1]Henry Ford Health, Detroit, MI, USA
[2]Michigan State University, East Lansing, MI, USA
Correspondence: szaidi7@hfhs.org

Abstract

Standard coherence metrics for biomedical topic models encode no clinical knowledge and cannot detect clinically implausible topic groupings. We propose SNOMED CT Wu–Palmer hierarchy distance as a post hoc, ontology-grounded diagnostic. On vascular surgery (47,318 articles) and craniofacial surgery (27,493 articles) corpora, the metric flags clinically heterogeneous topics that coherence misses—e.g., abdominal aortic aneurysm repair grouped with deep vein thrombosis ($d = 0.600$). Diagnostic signals are nearly identical across eight BERTopic embedding strategies including ontology-enhanced models, but diverge across model families: BERTopic alone produces a positive within- vs. cross-topic Cohen's d, while LDA, NMF, and Top2Vec at matched topic counts score below their own cross-topic baselines (Cohen's $d < 0$; Mann–Whitney $p > 0.99$). The score is therefore sensitive to topic-model output choice, not only to embedding choice within a single pipeline. A pre-clustering screening experiment finds near-zero correlation ($|\rho| < 0.08$) between embedding cosine and SNOMED CT similarity, arguing that ontological validation belongs after clustering rather than as an embedding screen. We additionally describe a two-stage UMLS-CUI stopword filter that preserves high-frequency domain-specific concepts which naive frequency filtering would discard. After one-time concept curation, the diagnostic itself is automated and requires no per-topic expert scoring.

1 Introduction

Topic models have become a standard tool for organizing and exploring biomedical literature at scale (Yeganova et al., 2018). Latent Dirichlet Allocation (Blei et al., 2003) and, more recently, BERTopic (Grootendorst, 2022) are widely applied in bibliometric analyses across clinical specialties, enabling researchers to identify thematic clusters, track temporal trends, and map the intellectual landscape of a field.

Evaluation of topic model quality relies primarily on intrinsic coherence metrics such as u_{mass} and C_v, which measure word co-occurrence statistics within discovered topics (Röder et al., 2015). While these metrics capture statistical regularity, they encode no domain knowledge. As a concrete example, a topic that groups articles on abdominal aortic aneurysm (AAA) repair—an elective *arterial reconstructive procedure*—with articles on deep vein thrombosis (DVT)—a *venous thromboembolic disease*—may achieve acceptable coherence scores because both share vascular vocabulary ("treatment," "graft," "outcome"). Yet the two occupy distinct branches of every clinical taxonomy, and merging them would mislead any researcher relying on the topic structure.

This gap between statistical coherence and clinical coherence has practical consequences. Researchers using topic models for systematic review screening (Thomas et al., 2011), research gap identification, or funding landscape analysis may draw conclusions from topics that conflate clinically distinct entities. A researcher asking "how has vascular surgery research evolved?" using a model that merges AAA repair and DVT into one topic is analyzing an artifact, not a literature structure.

The alternative—human expert evaluation—does not scale. Chang et al. (2009) demonstrated that human topic interpretability judgments are subjective and poorly correlated with statistical metrics. Hoyle et al. (2021) showed that even expert coherence ratings vary substantially, concluding that human evaluations have been largely abandoned by topic model developers. For a vascular surgery corpus producing 37 topics across 40 mapped clinical concepts, exhaustive pairwise clinical review requires evaluating $\binom{40}{2} = 780$ concept pairs—an effort no clinical reviewer can perform systematically, and one where reviewer differences

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 342–352

in training, subspecialty, and clinical culture would introduce arbitrary variation (Lau et al., 2014). An automated, ontology-grounded method is needed.

We address this gap by introducing SNOMED CT hierarchy distance as a post hoc, ontology-grounded diagnostic for biomedical topic models. SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) is the most comprehensive clinical terminology, organizing over 350,000 active concepts in a directed acyclic graph via `|is-a|` relationships (Pedersen et al., 2007; Melton et al., 2006). We apply Wu–Palmer normalized distance (Wu and Palmer, 1994) based on the Lowest Common Ancestor (LCA) to score the ontological coherence of topics produced by BERTopic-style pipelines (neural embeddings followed by UMAP+HDBSCAN clustering), and additionally compare against LDA, NMF, and Top2Vec under matched topic counts to test whether the diagnostic generalizes across topic-model families (§4.6). Contextualized topic models remain future work.

Our contributions are:

1. **Ontology-grounded post hoc diagnostic for biomedical topic models.** We define an SNOMED CT Wu–Palmer distance metric over topic-model outputs and show, on vascular (47,318 articles) and craniofacial (27,493 articles) corpora, that it flags clinically heterogeneous topics that coherence scores do not.

2. **Sensitivity across topic-model families.** The score produces near-identical signals across eight BERTopic embedding strategies (mxbai, PubMedBERT, SapBERT, BioLORD, cui2vec, MeSH TF-IDF, CUI co-occurrence, and an ensemble) but diverges across model families: BERTopic produces the only positive within- vs cross-topic Cohen's d on both corpora, while LDA, NMF, and Top2Vec at matched topic counts all score below their own cross-topic baselines (Cohen's $d < 0$, Mann–Whitney $p > 0.99$). This shows the metric is sensitive to topic-model output choice and not only to embedding choice within a single pipeline.

3. **Negative embedding screening result.** Across ten model–corpus combinations the cosine geometry of biomedical embeddings is only weakly correlated with SNOMED CT distance (Spearman $|\rho| < 0.08$), arguing that ontological validation is best applied after clustering rather than as an embedding screen—even for models trained on UMLS relations.

4. **Two-stage CUI stopword filtering.** We identify generic UMLS concepts (e.g., *Patients*, *Clinical Research*) that inflate CUI-based topic overlap scores and cascade-merge BERTopic outputs, and propose a frequency-plus-semantic-type filter that preserves high-frequency domain-specific concepts (e.g., *Cleft Palate* 19.1%, *Complication* 22.5%) which naive frequency filtering would incorrectly drop.

2 Background and Related Work

SNOMED CT and UMLS. SNOMED CT (Donnelly, 2006) is a clinical terminology of ∼350,000 active concepts organized as a polyhierarchical DAG via `|is-a|` relations; each concept has a stable SNOMED Concept Identifier (SCTID). The Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) aggregates over 200 vocabularies, assigns each concept a Concept Unique Identifier (CUI), and provides semantic types (e.g., T046 Pathologic Function) and crosswalks to MeSH, ICD, and SNOMED CT. In this paper *SCTID* refers to nodes used to compute hierarchy distance, and *CUI* refers to concepts extracted from text via SciSpacy (Neumann et al., 2019).

Topic models for biomedical literature. LDA (Blei et al., 2003) and its biomedical applications (Yeganova et al., 2018; Thomas et al., 2011) preceded the contemporary embedding-plus-clustering family typified by BERTopic (Grootendorst, 2022): encode with a language model, reduce via UMAP (McInnes et al., 2018), cluster with HDBSCAN. Domain-pretrained encoders such as PubMedBERT (Gu et al., 2021), SapBERT (Liu et al., 2021), BioLORD (Remy et al., 2024), and CUI-aware spaces such as cui2vec (Beam et al., 2020) are widely substituted at the embedding stage. Our primary experiments use this family; we additionally compare against LDA, NMF, and Top2Vec (§4.6) to test whether the diagnostic generalizes beyond embedding-based clustering. Contextualized topic models remain future work (§7). Topic quality is typically assessed via word-co-

occurrence coherence (Röder et al., 2015), which Chang et al. (2009) and Hoyle et al. (2021) show agrees poorly with human judgement, motivating automated, knowledge-grounded alternatives (Lau et al., 2014).

Ontology-grounded similarity. A long line of work uses hierarchical taxonomies to score concept similarity—Wu–Palmer (Wu and Palmer, 1994), Resnik, Lin, Jiang–Conrath—surveyed in Harispe et al. (2015) and applied to biomedical ontologies in Pedersen et al. (2007). SNOMED CT distance has been used for patient similarity (Melton et al., 2006). Bhattacharya et al. (2018) applied topic modeling *to* SNOMED codes in EHR data; we work in the opposite direction. The distinction between distributional *relatedness* and taxonomic *similarity* is established in Hill et al. (2015), and Mao and Fung (2020) showed that UMLS relation types are not uniformly captured by biomedical embeddings; our negative screening result (§4.4) extends these findings to SNOMED CT hierarchy distance specifically. To our knowledge, the present work is the first to apply SNOMED CT distance as a post hoc diagnostic on topic-model output of biomedical literature.

3 Methods

3.1 SNOMED CT Hierarchy Distance

We score topic-level ontological coherence using Wu–Palmer normalized distance (Wu and Palmer, 1994) over the SNOMED CT `|is-a|` hierarchy. Among the canonical taxonomy-based measures (Harispe et al., 2015), Wu–Palmer is depth-normalized (yielding $d \in [0, 1]$ for direct comparison across SNOMED branches of differing depth), does not require corpus-derived information content (unlike Resnik, Lin, and Jiang–Conrath, which would introduce a second corpus-dependent quantity into a metric meant to be *purely* ontological), and is the form most commonly applied to SNOMED CT in prior work (Pedersen et al., 2007).

For two concepts c_1 and c_2 with Lowest Common Ancestor (LCA) in the `|is-a|` hierarchy, Wu–Palmer normalized distance is

$$d(c_1, c_2) = 1 - \frac{2\,\mathrm{depth}(\mathrm{LCA})}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)},$$

where depth is the shortest path from the SNOMED CT root (Pedersen et al., 2007). The distance is 0 for identical concepts and approaches 1 for pairs sharing only a high-level ancestor. For polyhierarchical concepts with multiple `|is-a|` parents we use the shortest path, following standard practice (Harispe et al., 2015). We chose SNOMED CT over MeSH (another common UMLS-linked taxonomy) because SNOMED CT provides finer-grained subspecialty distinctions for surgical concepts and the procedural–disorder branch split that drives several of our representative failure cases; MeSH descriptors are nonetheless used in the embedding-comparison arm (Table 2, MeSH TF-IDF).

3.2 Concept-to-Article Mapping

For each specialty, we curate SNOMED CT concepts representing the major clinical domains: 75 concepts for vascular surgery (spanning aortic, carotid, peripheral arterial, venous, and access procedures and their corresponding disorders) and 48 concepts for craniofacial surgery (spanning congenital anomalies, facial trauma, reconstruction, tumors, and nerve injury). The asymmetry reflects vascular surgery's broader anatomical scope and broader publication remit: beyond procedure and disorder concepts spanning each subspecialty, the vascular set also covers pharmacotherapy, imaging, basic-science, and professional topics that the craniofacial registry, anchored more tightly to anatomy and pathology, largely omits. Each concept is associated with discriminating keywords (e.g., "carotid endarterectomy" → SCTID 174776006). An article maps to a concept by keyword search over both its title and abstract text—abstracts having been shown to carry the highest information density of any article section (Schuemie et al., 2004). Up to 100 articles per concept are retained.

Topic assignments from BERTopic provide the link between articles and topics, enabling determination of which SNOMED CT concepts are represented within each topic.

Coverage. Keyword-based mapping against both title and abstract text achieves 78.8% coverage of vascular articles ($n = 37{,}306$ of 47,318) and 61.0% of craniofacial articles ($n = 16{,}763$ of 27,493). All 37 vascular topics (100%) and 12 of 29 craniofacial topics (41.4%) contain $\geq$2 mapped concepts and are therefore evaluable. The remaining unmapped articles lack concept-specific keywords across both title and abstract text. Prior work found information density highest in abstracts, although full text provides broader coverage (Schuemie et al., 2004). MeSH-to-SNOMED crosswalks (Bodenrei-

der, 2004) could further improve coverage.

3.3 Topic-Level Ontological Validation

For each topic t containing mapped concepts $C_t = \{c_1, c_2, \ldots, c_k\}$ where $k \geq 2$, we compute the mean pairwise ontological distance:

$$\bar{d}(t) = \frac{2}{k(k-1)} \sum_{i<j} d(c_i, c_j) \qquad (1)$$

Using a heuristic default threshold of $\bar{d}(t) > 0.5$, topics are flagged as *clinically heterogeneous*. We additionally identify *illustrative cross-domain pairs*—specific concept pairs within a topic that make the heterogeneity clinically interpretable.

3.4 Embedding Screening (Stage 0)

We test whether embedding geometry mirrors clinical ontology, which would allow embedding selection to serve as an upstream proxy. For each model, we compute centroid embeddings per SNOMED CT concept by averaging embeddings of mapped articles, compute pairwise cosine similarities, and correlate with SNOMED CT Wu–Palmer similarity using Spearman's ρ.

3.5 CUI Stopwords for Topic Reconciliation

When using UMLS Concept Unique Identifiers (CUIs) extracted from title and abstract text via SciSpacy (Neumann et al., 2019) to compare topics—e.g., via Jaccard overlap of per-topic CUI profiles—generic medical concepts that appear across most articles inflate similarity scores and cause spurious cascade merges. We identify these "medical CUI stopwords" empirically via a two-stage procedure. First, for each CUI extracted from the corpus, we compute document frequency (the fraction of articles containing the concept). CUIs exceeding a 15% prevalence threshold are candidates for filtering. Second, we classify each candidate by its UMLS semantic type: CUIs belonging to generic types (T079 Temporal Concept, T078 Idea or Concept, T080 Qualitative Concept, T081 Quantitative Concept, T101 Patient or Disabled Group, T062 Research Activity, etc.) are designated stopwords, while CUIs belonging to clinical types (T046 Pathologic Function, T061 Therapeutic Procedure, T047 Disease or Syndrome, T023 Body Part, etc.) are retained regardless of frequency. This two-stage approach prevents filtering of high-frequency domain-specific concepts that are substantively important within a specialty corpus.

On the vascular corpus (47,318 articles), the frequency threshold identifies 27 CUIs at $\geq$15%, of which 21 are generic (led by *Patients*, C0030705, 77.7%) and 6 are clinical (e.g., *Complication*, 22.5%; *Interventional procedure*, 18.8%). The semantic type filter retains the clinical CUIs and removes only the 21 generic stopwords. Without filtering, CUI-based topic overlap scores are dominated by ubiquitous concepts; in early experiments, Jaccard-based merge rules cascaded 37 topics down to 2. After filtering, no spurious merges occur (Section 4.2).

3.6 Post-Modeling Ontology Reconciliation

As a complementary step, we apply a two-pass pipeline. After BERTopic clustering (Pass 1), a CUI-based reconciliation pass (Pass 2) scores all topic pairs on Jaccard overlap of their CUI profiles (with stopwords removed) and merges pairs exceeding an empirically determined similarity threshold of 0.7 (requiring substantial CUI overlap before merging), with an anti-cascade constraint (each topic merges at most once per pass). This tests whether statistically discovered topics contain redundant clinical content that should be consolidated.

3.7 Corpora

Vascular surgery. 47,318 articles from 8 vascular surgery journals,[1] spanning 1984–2026. Journals were identified via PubMed's NLM Catalog under the "Vascular Surgery" and "Vascular Diseases" Broad Subject Terms, selecting all journals whose primary scope is vascular surgical practice. To test metric robustness across upstream modeling choices, we apply SNOMED CT validation to topic model outputs from eight embedding strategies, selected to span the major design axes available for biomedical text representation: general-purpose transformer (mxbai-embed-large-v1, 335M parameters (Lee et al., 2024)), biomedical transformers (PubMedBERT, 110M (Gu et al., 2021); SapBERT, 110M (Liu et al., 2021)), ontology-grounded models (BioLORD-2023, 110M (Remy et al., 2024); cui2vec, 500-d (Beam et al., 2020); Poincaré embeddings of SNOMED CT `|is-a|` edges, 200-d), a CUI co-occurrence TF-IDF baseline (200-d via SVD), and an ensemble (L2-normalized average of PubMedBERT + BioLORD). All use

[1] JVS, Annals of Vascular Surgery, EJVES, J. Endovascular Therapy, Vascular, JVS-VLD, Seminars in Vascular Surgery, JVS-Vascular Science.

Model	All-pairs	Same-dom.	Weighted	Single.
mxbai	0.711	0.394	0.711	10
PubMedBERT	0.722	0.367	0.742	1
BioLORD	0.715	**0.337**	0.735	2
Ensemble	**0.701**	0.347	**0.711**	**0**

Table 1: SNOMED CT ontological validation applied to four embedding models on JVS ($n = 14{,}290$). The metric yields similar diagnostic signals ($\bar{d} \approx 0.70$–0.74) across models, while still revealing model-specific failure modes (e.g., mxbai produces 10 singleton concepts vs. 0 for ensemble).

BERTopic with UMAP + HDBSCAN clustering (McInnes et al., 2018) (`min_cluster_size` = 300, `min_samples` = 15).

Craniofacial surgery. 27,493 articles from 9 journals,[2] spanning 1994–2025, identified via PubMed NLM Catalog under craniofacial and oral surgery subject headings. BERTopic with mxbai-embed-large-v1 embeddings, producing 29 topics.

3.8 Classical Topic-Model Baselines

To test whether the SNOMED CT diagnostic discriminates across topic-model families (not only embedding choices within BERTopic), we additionally fit three classical models on both corpora with matching topic counts ($k = 37$ vascular, $k = 29$ craniofacial): LDA (Blei et al., 2003) over a bag-of-words representation (min_df = 5, max_df = 0.95), NMF over TF-IDF features (default sklearn settings), and Top2Vec, which jointly learns document and word embeddings followed by HDBSCAN clustering and hierarchical reduction to the target k. The same SNOMED concept mapping and Wu–Palmer scoring pipeline is then applied to each model's topic assignments.

4 Results

4.1 Vascular Surgery Corpus

To assess whether the metric produces stable diagnostic signals across different upstream embedding choices, we apply it to four embedding models on the single-journal JVS corpus (Table 1; 14,290 articles, 75 SNOMED concepts) and then extend to eight embedding strategies on the full multi-journal corpus (47,318 articles, 8 journals).

With under-constrained clustering (`min_cluster_size` = 40), the mxbai model produces a mega-cluster (Topic 0, 36% of corpus) spanning 4+ SNOMED branches with mean intra-topic distance = 0.768 across 78 concept pairs. Six illustrative cross-domain pairs show clinically implausible groupings within this single topic:

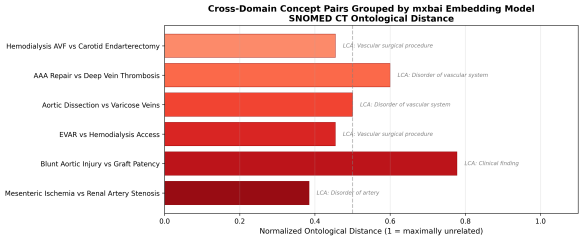

Figure 1: Six representative cross-domain concept pairs grouped within a single topic by the mxbai model. Each bar shows the Wu–Palmer SNOMED CT distance between clinically unrelated concepts that the topic model erroneously merged. Standard coherence metrics do not flag this topic.

- Hemodialysis AVF ↔ Carotid endarterectomy ($d = 0.455$; LCA: Vascular surgical procedure)
- AAA repair ↔ Deep vein thrombosis ($d = 0.600$; LCA: Disorder of vascular system)
- Aortic dissection ↔ Varicose veins ($d = 0.500$; LCA: Disorder of vascular system)
- Blunt aortic injury ↔ Graft patency ($d = 0.778$; LCA: Clinical finding)
- EVAR ↔ Saphenous vein ablation ($d = 0.500$; LCA: Endovascular procedure)
- PAD claudication ↔ IVC filter ($d = 0.714$; LCA: Disorder)

In this setting, the metric flags failures that coherence scores do not: the mega-cluster achieves acceptable coherence despite grouping concepts from 4+ SNOMED branches. Across all four models, the metric flags clinically heterogeneous topics (all-pairs $\bar{d}$ range: 0.701–0.722), suggesting that the diagnostic signal is not confined to a single embedding choice.

Multi-journal robustness (8 embedding strategies). On the full 47,318-article, 8-journal vascular corpus with appropriately constrained clustering (`min_cluster_size` = 300), we test eight embedding strategies (Table 2). Topic counts range

[2] J. Craniofacial Surgery, J. OMS, Cleft Palate–Craniofacial J., Int. J. OMS, J. CMF Surgery, PRS, Br. J. OMS, Oral Surg. Oral Med. Oral Pathol., Head & Face Medicine.

Embedding	K	Out.%	Sil.	C_v
Ensemble	**37**	**18.3**	0.526	**0.792**
SapBERT	33	23.9	0.465	0.757
SapBERT+Snomed2Vec	33	23.9	0.464	0.757
CUI co-occurrence	49	28.9	0.528	0.718
MeSH TF-IDF	37	22.3	0.597	0.639
SapBERT+cui2vec	20	26.6	0.391	0.796
cui2vec	5	6.7	0.170	0.392
Snomed2Vec	26	9.3	−0.065	—

Table 2: Eight embedding strategies on the vascular corpus ($n = 47{,}318$, 8 journals). K = topics, Out.% = outlier rate, Sil. = silhouette. Ontology-enhanced embeddings (Snomed2Vec, cui2vec) do not outperform the baseline ensemble; Snomed2Vec produces negative silhouette, indicating cluster overlap. Bold = recommended model.

CUI (Name)	Doc Freq	%
C0030705 (Patients)	36,752	77.7
C0008972 (Clinical Research)	18,602	39.3
C0439234 (year)	16,739	35.4
C0868928 (Case)	12,990	27.5
C0039798 (therapeutic)	12,360	26.1
C0332281 (Associated with)	12,278	25.9
C0220825 (Evaluation)	11,707	24.7
…	(27 total $\geq 15\%$)	

Table 3: Top generic UMLS CUIs by document frequency in the vascular corpus. These "medical stopwords" must be filtered before CUI-based topic comparison to prevent cascade merges.

from 5 (cui2vec, severe under-clustering) to 49 (CUI co-occurrence). The recommended model (PubMedBERT + BioLORD ensemble, 37 topics, $C_v = 0.792$, 18.3% outliers) eliminates the mega-cluster problem observed with under-constrained parameters, producing topics that domain experts can inspect. Across the six viable arms, mean within-topic distances range from 0.767 to 0.770 (cross-topic baseline: 0.771), with Cohen's d between 0.003 and 0.015. This consistency suggests the findings are not artifacts of a single embedding model. The small positive Cohen's d reflects vascular topics spanning both procedures and disorders within each domain.

Clustering parameters vs. embedding choice. The jump from `min_cluster_size` $= 40$ (43 topics, 30.2% outliers, mega-clusters) to `min_cluster_size` $= 300$ (37 topics, 18.3% outliers, no mega-clusters) suggests that HDBSCAN parameter tuning has a larger effect on clinical face validity than embedding model selection. A topic count in the 30–50 range provided the clearest balance of subspecialty granularity in our data, while under-constrained clustering ($K > 100$) or over-constrained clustering ($K < 20$) produced topics that were implausibly coarse or fragmented.

4.2 CUI Stopwords

Table 3 shows the ten most prevalent CUIs in the vascular corpus. Without filtering, CUI Jaccard overlap between topic pairs is dominated by these generic concepts: in early experiments without a stoplist, Pass 2 reconciliation cascade-merged 37 topics to 2. After filtering the 21 generic CUIs (retaining 6 clinical CUIs via semantic type classification), Pass 2 produces zero merges in our experiments across all eight embedding arms—suggesting that appropriately constrained HDBSCAN already produces non-redundant topics.

Cross-corpus validation. To test whether the same generic-term pattern appears in a second corpus, we performed an analogous analysis on the craniofacial corpus using MeSH descriptor document frequencies (27,305 articles). Of 12 MeSH terms at $\geq$15%, 9 are generic (Humans 93.4%, Male 58.7%, Female 57.5%, Adult 36.7%, Retrospective Studies 25.4%) and 3 are clinical (Mandible 19.3%, Cleft Palate 19.1%, Cleft Lip 16.4%). Frequency-only filtering would incorrectly remove *Cleft Palate*—a defining condition of the specialty appearing in one-fifth of articles—while semantic type classification correctly retains it. The same structural pattern (generic demographics and study-design terms dominating, with domain-specific clinical terms at the boundary) appears in both corpora, supporting the two-stage approach as reusable across these two specialties.

4.3 Craniofacial Surgery Corpus

Table 4 presents per-topic ontological distances for the 12 craniofacial topics containing $\geq$2 mapped SNOMED concepts (48 concepts, 61 intra-topic pairs).

Nine topics ($\bar{d} < 0.5$) are ontologically coherent. T8 (Facial trauma) groups mandible, orbital, zygoma, nasal, and Le Fort fractures—all under "Fracture of bone of face" in SNOMED CT ($\bar{d} = 0.136$). T3 (Cleft lip and palate) groups cleft lip, cleft palate, combined cleft, and Pierre Robin sequence ($\bar{d} = 0.220$).

Three topics ($\bar{d} \geq 0.5$) are flagged as heterogeneous. Representative examples include:

- Rhinoplasty $\leftrightarrow$ Mandible fracture ($d = 1.000$;

Topic	$\bar{d}$	k	Pairs
T2: Orthognathic surgery	0.116	3	3
T8: Facial trauma	0.136	5	10
T1: Bone / distraction	0.200	2	1
T3: Cleft lip and palate	0.220	4	6
T7: Nerve injury	0.278	3	3
T13: Fillers / fat grafting	0.333	2	1
T20: Beauty / aesthetics	0.333	2	1
T6: Tumors and cysts	0.359	3	3
T0: Craniosynostosis	0.434	8	28
T12: Condylar / TMJ	0.511	3	3
T4: Flap reconstruction	0.600	2	1
T16: Hemangioma / vasc. malf.	0.600	2	1

Table 4: Per-topic SNOMED CT distances for craniofacial surgery ($n = 27{,}493$, mxbai embeddings). The midrule separates topics below the heuristic threshold ($\bar{d} < 0.5$) from those at or above it ($\bar{d} \geq 0.5$). k = number of mapped SNOMED concepts.

LCA: SNOMED CT root — elective aesthetic procedure vs. acute trauma)

- Cleft lip ↔ Ameloblastoma ($d = 0.667$; LCA: Disease — congenital anomaly vs. neoplasm)
- Craniosynostosis ↔ Burn injury ($d = 0.556$; LCA: Disease — congenital vs. traumatic)
- Facial nerve palsy ↔ Osteonecrosis of jaw ($d = 0.636$; LCA: Disease — neurological vs. bone pathology)

This thresholding provides a simple operational mechanism for topic-level filtering, enabling identification of topics that may require splitting or further inspection. The overall mean of per-topic mean distances is 0.343 (SD = 0.159), compared to 0.768 for the mxbai vascular mega-cluster. The craniofacial corpus shows better topic-level ontological coherence overall, likely because the 29-topic structure provides finer granularity than the mxbai vascular run's coarser clustering.

Threshold sensitivity. We adopt $\bar{d} > 0.5$ as a heuristic default separating single-domain topics from cross-domain groupings. Discriminative power varies by specialty and granularity—vascular topics inherently bridge procedure and disorder branches—and specialty-specific calibration is recommended.

4.4 Embedding Screening: A Negative Result

Table 5 presents Spearman correlations between embedding cosine similarity and SNOMED CT Wu–Palmer similarity.

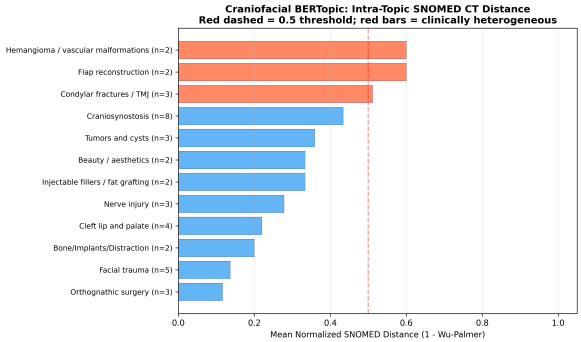

Figure 2: Per-topic SNOMED CT ontological distances for the craniofacial corpus. Topics below the dashed threshold ($\bar{d} = 0.5$) are ontologically coherent; topics above are flagged as clinically heterogeneous.

Model	Corpus	ρ	p	Pairs
mxbai	Vasc. (JVS)	0.037	0.305	780
PubMedBERT	Vasc. (JVS)	0.065	0.071	780
BioLORD	Vasc. (JVS)	0.013	0.713	780
Ensemble	Vasc. (JVS)	0.045	0.211	780
Ensemble	Vasc. (8J)	0.005	0.900	780
SapBERT	Vasc. (8J)	0.004	0.912	780
CUI co-occ.	Vasc. (8J)	0.023	0.522	780
MeSH TF-IDF	Vasc. (8J)	0.043	0.234	780
SapBERT+cui2vec	Vasc. (8J)	0.025	0.493	780
mxbai	Craniofacial	0.022	0.468	1,128

Table 5: Embedding screening: Spearman ρ between cosine similarity and SNOMED CT Wu–Palmer similarity across 10 model–corpus combinations. All show $|\rho| < 0.08$. Vasc. (JVS) = single-journal; Vasc. (8J) = 8-journal corpus.

All ten model–corpus combinations show near-zero correlation ($|\rho| < 0.08$). Most comparisons come from the vascular corpus, but the same near-zero pattern also appears in the craniofacial mxbai run. Within vascular data, the result is consistent across general-purpose (mxbai), biomedical (PubMedBERT, SapBERT), ontology-grounded (BioLORD, CUI co-occurrence, MeSH TF-IDF), and hybrid models, and across both single-journal and multi-journal settings. Notably, BioLORD-2023—which was explicitly trained on UMLS concept relationships including SNOMED CT (Remy et al., 2024)—shows $\rho = 0.013$ ($p = 0.71$).

Embedding models organize biomedical text by literature co-occurrence patterns, not by clinical ontological relationships (Hill et al., 2015). Concepts that are ontologically distant (e.g., AAA and DVT: different vascular systems, different pathophysiology) co-occur in the same journals, the same patient populations, and the same review articles. The embedding space reflects this co-occurrence

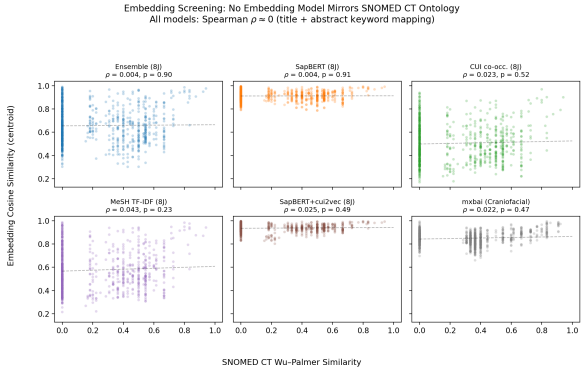

Figure 3: Embedding screening: cosine similarity versus SNOMED CT Wu–Palmer similarity for all concept pairs. All ten model–corpus combinations show Spearman $\rho \approx 0$, suggesting that embedding geometry does not mirror clinical ontological structure.

structure—distributional *relatedness*—not the clinical taxonomy—ontological *similarity* (Hill et al., 2015).

4.5 Informal Clinician Face-Validity Check

We include an informal three-surgeon assessment of one pipeline arm to illustrate—not to validate—the diagnostic's qualitative behaviour. Two vascular surgeons and one acute care surgeon *jointly* discussed the 37 topics from Arm A (ensemble) and proposed merge and split groupings; ratings were not collected independently, no blinding was applied, and no formal scoring rubric was used. We then compared this shared assessment with a domain-level split audit that classifies each topic by its distribution across nine vascular anatomical domains (Aortic, Cerebrovascular, Peripheral, Venous, Access, Visceral, Renal, Trauma, Thoracic Outlet). This comparison only covers Arm A on the vascular corpus, 16 judgements, and one specialty; it should not be read as evidence that the metric tracks clinical judgement in general (§7).

On these 16 judgements, the metric agreed with the surgeons on 4 merge groups comprising 13 topics in 12/13 cases (92%; e.g., Topics 7, 15, 23, 35—aortic aneurysm repair—scored zero heterogeneity with >83% Aortic concentration) and on 2 of 3 split requests (67%); the one miss separated amputation from other peripheral arterial topics, an intra-domain distinction that domain-level analysis cannot capture. Combined directional agreement was 14/16 (88%). We report this number for context only; it cannot establish metric validity. A blinded, independently rated, multi-specialty clinician study with inter-rater agreement is needed and is left to future work.

Corpus	Method	Eval.	$\bar{d}_w$	$\bar{d}_c$	Cohen's d
Cranio.	BERTopic	12/29	0.343	0.768	+0.159
	LDA	28/29	0.715	0.703	−0.043
	NMF	29/29	0.726	0.713	−0.046
	Top2Vec	29/29	0.726	0.712	−0.050
Vasc.	BERTopic	37/37	0.767	0.771	+0.015
	LDA	36/37	0.767	0.747	−0.073
	NMF	37/37	0.770	0.750	−0.071
	Top2Vec	37/37	0.769	0.749	−0.071

Table 6: SNOMED diagnostic applied to classical and neural topic models. $\bar{d}_w$ = mean within-topic distance; $\bar{d}_c$ = cross-topic baseline; Eval. = topics with $\geq$2 mapped concepts. Positive Cohen's d indicates within-topic < cross-topic (clinically coherent grouping). All three alternatives show negative Cohen's d on both corpora (Mann–Whitney $p > 0.99$).

4.6 Classical Baseline Comparison

To test whether the diagnostic generalizes beyond BERTopic-style pipelines, we apply it to three alternative topic models—Latent Dirichlet Allocation (LDA; Blei et al. 2003), Non-negative Matrix Factorization (NMF), and Top2Vec—run on both corpora with matching topic counts ($k = 37$ vascular, $k = 29$ craniofacial). LDA uses bag-of-words input, NMF uses TF-IDF, and Top2Vec jointly learns document and word embeddings with hierarchical topic reduction.

Table 6 shows that all three classical alternatives score below their own cross-topic baselines on both corpora, while BERTopic alone produces positive within- vs. cross-topic separation. On the craniofacial corpus, BERTopic achieves $\bar{d}_w = 0.343$ (well below the cross-topic baseline of 0.768), while LDA, NMF, and Top2Vec produce $\bar{d}_w \approx 0.72$—statistically indistinguishable from their own cross-topic baselines (Mann–Whitney $p > 0.99$; Cohen's $d < 0$). LDA generates 10 mega-clusters containing >40 mapped concepts each (of 48 total), indicating that most concepts co-occur in nearly every topic; NMF and Top2Vec produce 6 and 3 mega-clusters respectively. On the vascular corpus, all four methods show high $\bar{d}_w \approx 0.77$, but only BERTopic produces a (small) positive Cohen's d.

5 Discussion

What the metric does, and what it does not do. The diagnostic flags topics whose mapped SNOMED CT concepts span distant hierarchy branches—a failure mode that coherence metrics, which score word co-occurrence rather than on-

tological position, do not detect. We frame the metric as a post hoc complement to coherence, not a replacement: a topic may be coherent yet ontologically heterogeneous, and vice versa. After one-time concept curation, the scoring step is automated, but curation, threshold selection, and the interpretation of high-distance topics still require domain judgement.

The co-occurrence–ontology gap. The near-zero screening correlation ($|\rho| < 0.08$ across ten model–corpus combinations) extends Mao and Fung (2020) to the specific case of SNOMED CT hierarchy distance: even models trained on UMLS relations (BioLORD-2023, SapBERT) produce embedding spaces that capture distributional relatedness, not taxonomic position. The practical implication is that embedding choice cannot stand in for ontological validation, and that ontological scoring is best applied after clustering.

Clustering constraints matter more than embedding choice. Our eight-arm comparison suggests HDBSCAN parameter selection matters more for clinical face validity than embedding choice within the BERTopic family. Under-constrained clustering (`min_cluster_size` $= 40$) produces mega-clusters spanning 4+ SNOMED branches, whereas over-constrained settings ($K < 20$) collapse subspecialty distinctions. In our data, 30–50 topics gave the best balance, so future studies should report HDBSCAN sensitivity alongside embedding selection.

Sensitivity across topic-model families. The diagnostic is a post hoc score, not a constraint on the underlying models. What the cross-family comparison shows is that the score is sensitive to topic-model output choice, not only to embedding choice within a single pipeline (Table 2 vs. Table 6). We do not claim that BERTopic is universally superior to LDA, NMF, or Top2Vec; we observe only that its outputs have higher SNOMED CT ontological coherence on these two corpora at the tested topic counts. The cross-family direction is consistent across both corpora, but the magnitude of BERTopic's advantage is much larger on craniofacial ($+0.159$) than on vascular ($+0.015$).

Medical CUI stopwords. Generic UMLS concepts dominate CUI-based topic overlap scores in our data; any system using CUI profiles for clustering or merge decisions will encounter the same noise floor. The two-stage filter (frequency $\geq$15% followed by UMLS semantic type) preserves high-frequency domain-specific terms (*Cleft Palate* 19.1%, *Complication* 22.5%) that naive frequency thresholding would discard, while removing demographic, study-design, and temporal qualifiers. The generic/substantive boundary remains context-dependent; semantic types provide a principled default with room for domain-specific override. Pass 2 ontology reconciliation produced zero merges across all eight arms, indicating that constrained HDBSCAN already discovers non-redundant clinical topics in our data.

6 Conclusion

SNOMED CT Wu–Palmer distance flags clinically heterogeneous biomedical topics that word-coherence metrics do not, and it is sensitive to topic-model output choice across families. Across two surgical corpora, eight BERTopic embedding strategies, and the LDA, NMF, and Top2Vec baselines, only BERTopic produces positive within- vs. cross-topic separation (Tables 2, 6). The near-zero embedding-screening correlation argues that ontological validation belongs *after* clustering, not as an embedding screen. We additionally contribute a two-stage CUI stopword filter and find that HDBSCAN constraints may matter more for clinical face validity than embedding choice. Contextualized topic models and a blinded independently rated clinician study remain future work (§7).

7 Limitations

Scope of topic-model family. The primary experiments use BERTopic with UMAP+HDBSCAN. A preliminary comparison with LDA, NMF, and Top2Vec (§4.6) shows the diagnostic generalizes across topic-model families, but contextualized topic models, MeSH-guided variants, and systematic hyperparameter sweeps remain untested. The screening result (§4.4) specifically pertains to embedding-based pipelines.

Concept curation and coverage. Concept curation requires domain knowledge, though the effort is modest (∼50–75 concepts per specialty). Keyword mapping against title and abstract text covers 78.8% of vascular and 61.0% of craniofacial articles; full-text coverage would be broader (Schuemie et al., 2004), and MeSH-to-SNOMED crosswalks via UMLS (Bodenreider, 2004) could improve recall. Topics with fewer than two mapped concepts cannot be evaluated.

Metric calibration. SNOMED CT branch depths vary, and the $\bar{d} > 0.5$ flagging threshold is heuristic and specialty-sensitive; SNOMED CT also does not always capture cross-cutting themes (complications, perioperative care, imaging findings, outcomes) the way clinicians conceptualize them. Elevated distance should prompt inspection rather than be treated as categorical error. The CUI stoplist should be re-derived per specialty.

Method assumptions. Several methodological choices warrant flagging. The article-to-concept mapping caps each concept at 100 articles, which may distort topic statistics for high-frequency concepts; topics with only two mapped concepts are flagged on a single pair, exaggerating noise. The embedding-screening centroid aggregation may wash out fine-grained substructure, and the Spearman correlation assumes a monotonic alignment between embedding cosine and SNOMED CT similarity that need not hold. Wu–Palmer depth normalization assumes comparability across SNOMED branches, which may not hold uniformly between procedural and disorder hierarchies. Both validated corpora are surgical specialties; generalization to non-procedural domains (e.g., psychiatry, primary care, genomics) is untested, and temporal-span differences between corpora (1984–2026 vs. 1994–2025) are not analyzed.

Clinician validation. The three-surgeon assessment (§4.5) was a single joint discussion on one pipeline arm in one specialty, with no independent scoring, blinding, or rubric. It cannot establish metric validity. A blinded, independently rated, multi-specialty study with inter-rater agreement, power calculation, and a formal rubric is needed before the metric should be cited as predictive of clinical judgement.

References

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. *Pacific Symposium on Biocomputing*, 25:295–306.

Moumita Bhattacharya, Claudine Jurkovitz, and Hagit Shatkay. 2018. Co-occurrence of medical conditions: Exposing patterns through probabilistic topic modeling of SNOMED codes. *Journal of Biomedical Informatics*, 82:31–40.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22, pages 288–296.

Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. ACL.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread – new fluffy embeddings model. Mixedbread AI Blog. `https://www.mixedbread.ai/blog/mxbai-embed-large-v1`.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. 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*, pages 4228–4238. ACL.

Yuqing Mao and Kin Wah Fung. 2020. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *Journal of the American Medical Informatics Association*, 27(10):1538–1546.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29):861.

Genevieve B. Melton, Sonia Parsons, Frances P. Morrison, Adam S. Rothschild, Marianthi Markatou, and George Hripcsak. 2006. Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*, 39(6):697–705.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. ACL.

Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM.

Martijn J. Schuemie, Marc Weeber, Bob J. A. Schijvenaars, Erik M. van Mulligen, Cornelis C. van der Eijk, Rob Jelier, Barend Mons, and Jan A. Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.

James Thomas, John McNaught, and Sophia Ananiadou. 2011. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138. ACL.

Lana Yeganova, Sun Kim, Grigory Balasanov, and W. John Wilbur. 2018. Discovering themes in biomedical literature using a projection-based algorithm. *BMC Bioinformatics*, 19:269.