Dr-BERT-NL at #SMM4H–HeaRD 2026: DOKTERBERT – Ontology-Grounded Contextual Representations for Dutch Clinical NLP

Gijs Danoe[1,2] **Andreas Voss**[1] **Axel Hamprecht**[2] **Matthijs S. Berends**[1,3]

[1]Department of Medical Microbiology and Infection Prevention,
University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[2]Institute of Medical Microbiology and Virology,
Carl von Ossietzky Universität, Oldenburg, Germany
[3]Department of Medical Epidemiology, Certe Foundation, Groningen, The Netherlands

g.danoe@umcg.nl, a.voss@umcg.nl, axel.hamprecht@uni-oldenburg.de, m.s.berends@umcg.nl

Abstract

Clinical language models are typically pre-trained with self-supervised objectives whose geometry reflects linguistic co-occurrence rather than clinical knowledge structure. For downstream tasks that operate directly on the representation space, without task-specific fine-tuning, this gap limits what the model can do. We introduce DOKTERBERT (Dutch Ontology-grounded Knowledge-injected Text Encoder for Representations using BERT), a Dutch clinical language model pre-trained with a structure-aware contrastive objective that aligns contextual span representations to SNOMED concept anchors, with negative pressure weighted by graph distance in the SNOMED hierarchy. We evaluate DOKTERBERT against three Dutch baselines (RobBERT, MedRoBERTa.nl, and MedRoBERTa.nl-SapBERT) through supervised named entity recognition on MultiClinNER-nl and an unsupervised representation analysis spanning retrieval, clustering, entity linking, and concept-level separation. On supervised NER, all four models perform comparably; on the representational evaluations, DOKTERBERT separates from every baseline. Standard fine-tuning evaluation obscures pre-training-level differences in representation quality that representation analysis exposes, and these differences matter for clinical applications that depend on embedding geometry.

1 Introduction

Clinical natural language processing increasingly relies on contextual language models, whose representations underpin a growing range of downstream tasks. These representations are learned through self-supervised objectives on clinical text, and their geometry reflects the statistical co-occurrence structure of that text rather than the structure of clinical knowledge itself. Two concepts that are clinically distinct may appear in systematically similar contexts; two that are clinically equivalent may co-occur rarely through discourse structure alone. The resulting representation space encodes linguistic distribution, not clinical meaning.

For tasks that use the geometry directly, without task-specific fine-tuning, the ability of the space to resolve fine-grained clinical distinctions largely determines what the model can do. Similarity-based retrieval, clustering, and anomaly detection over clinical text all depend on the representation space having clinical structure rather than only linguistic structure. These tasks operate on clinically informative spans extracted from noisy text; named entity recognition provides the filter that isolates those spans, making the quality of the underlying representation space consequential rather than hidden beneath the noise of full-document pooling.

We introduce DOKTERBERT (Dutch Ontology-grounded Knowledge-injected Text Encoder for Representations using BERT), a Dutch clinical language model pre-trained with a structure-aware contrastive objective that aligns contextual span representations to SNOMED CT (Donnelly, 2006) concept anchors. SNOMED CT is the international standard clinical terminology, organising clinical concepts into an IS-A hierarchy that supports semantic comparison between concepts. The contrastive signal is weighted by graph distance in the SNOMED hierarchy, concentrating discriminative pressure on semantically adjacent concepts. We evaluate on MultiClinNER-nl (Gallego-Donoso et al., 2026), the Dutch subtask of the MultiClinAI shared task at the SMM4H/HeaRD workshop (Lopez-Garcia et al., 2026), through supervised named entity recognition and an unsupervised representation analysis against three Dutch baselines.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 154–159

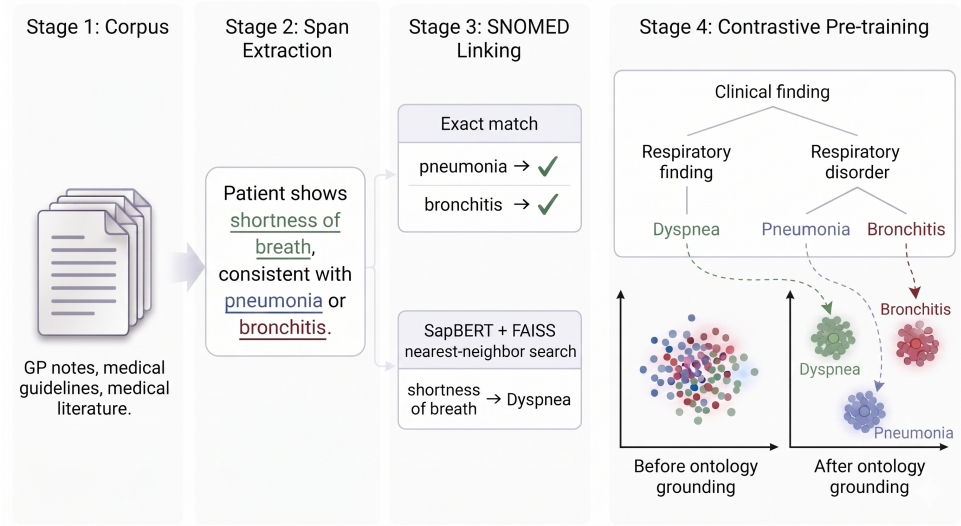

Figure 1: DOKTERBERT pre-training pipeline. Spans extracted from the corpus are linked to SNOMED concepts (stages 1–3) and aligned to concept anchors through distance-weighted contrastive pre-training (stage 4).

2 Related Work

The Dutch clinical NLP landscape is narrow. MedRoBERTa.nl (Verkijk and Vossen, 2021) adapts RoBERTa to Dutch clinical text through pre-training on hospital-based electronic health records, and is the backbone for downstream clinical NLP in Dutch. Its pre-training corpus does not include primary care text, which limits its coverage of the linguistic register and clinical vocabulary used in general practice. MedRoBERTa.nl has no ontology grounding: its geometry reflects the distribution of its pre-training corpus rather than the structure of clinical knowledge. Term-level ontology grounding is introduced by SapBERT (Liu et al., 2021), which contrastively aligns isolated concept names against UMLS; MedRoBERTa.nl-SapBERT (Hartendorp et al., 2024) extends this to Dutch clinical terminology by applying the SapBERT objective on top of MedRoBERTa.nl. Both operate on isolated terms rather than on spans in clinical context and were designed for entity linking rather than as general contextual encoders.

3 Method

DOKTERBERT is trained through the pipeline in Figure 1: candidate spans are extracted from a Dutch clinical corpus, linked to SNOMED concepts, and aligned to concept anchors through distance-weighted contrastive pre-training.

3.1 Pretraining Corpus and Initialization

DOKTERBERT is initialized from MedRoBERTa.nl and continues pre-training on 2.33 GB of Dutch clinical text: GP consultation notes from the AHON registry (Twickler et al., 2024), Nederlands Tijdschrift voor Geneeskunde (NTVG) articles, clinical and pharmacological guidelines, and patient-facing health information (Table 1). This extends MedRoBERTa.nl's hospital-based foundation into primary care. Reference and educational sources were included to broaden SNOMED concept coverage beyond routine GP consultations, addressing the long tail of clinical terminology the contrastive objective requires.

Source	Size (GB)
GP consultation notes	1.14
NTVG articles	0.91
NHG guidelines	0.13
Dutch medical Wikipedia	0.06
Apotheek.nl	0.05
Farmacotherapeutisch Kompas	0.02
MultiClinNER-nl (train)	0.01
UMCG patient information	0.004
RIVM	0.004
Total	2.33

Table 1: DOKTERBERT pre-training corpus by source.

3.2 Span Extraction

Candidate medical spans are extracted from the corpus using spaCy dependency parsing, retaining nouns, noun phrases, proper nouns, and stan-

dalone adjectives. This yields 69.8M candidate spans across the corpus.

3.3 SNOMED Linking

Each candidate span is linked to a SNOMED CT concept by exact string match against the Dutch SNOMED term set, with a SapBERT-based similarity fallback (cosine threshold 0.85) for unmatched spans. Spans below threshold are excluded from contrastive training.

This produces 11.4M linked spans covering 30,408 unique SNOMED concepts, with roughly half of linking from exact match and half from SapBERT similarity.

3.4 Contrastive Pre-training

Span and anchor representations. Span representations s are mean-pooled final-layer hidden states of the span tokens in their original sentence. Concept anchors a_c are the [CLS] embedding of each concept's preferred Dutch term under the base model, held fixed during training.

SNOMED graph distance. Concept distances $d(c_i, c_j)$ are shortest-path lengths in the SNOMED IS-A hierarchy, precomputed for all pairs in the training corpus.

Distance-weighted contrastive loss. Let $z(c) = \exp(\mathrm{sim}(s, a_c)/\tau)$ denote the exponentiated cosine similarity between span s and anchor a_c, where τ is a temperature hyperparameter. The InfoNCE loss with graph-distance-weighted negatives is

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{z(c^+)}{z(c^+) + \sum_{c^- \in \mathcal{N}} w(c^+, c^-)\, z(c^-)} \tag{1}$$

where $\mathcal{N}$ is the set of negative concepts drawn from other medical spans in the same batch, with c^+ excluded, and the weight function is

$$w(c^+, c^-) = \exp\left(-\frac{d(c^+, c^-)}{\sigma}\right). \tag{2}$$

The weight function concentrates contrastive pressure on semantically adjacent concepts.

Training objective. The contrastive loss is combined with standard masked language modeling loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLM}} + \alpha \cdot \mathcal{L}_{\text{InfoNCE}}. \tag{3}$$

Training details. We continue pre-training from MedRoBERTa.nl for one epoch with contrastive weight $\alpha = 0.2$, temperature $\tau = 0.07$, and graph-distance decay $\sigma = 15$. Full hyperparameters are listed in Appendix A.

4 Experiments and Results

4.1 Supervised Named Entity Recognition

Table 2 reports per-entity F1 on MultiClinNER-nl. All four models fall within a narrow band, with general-purpose RobBERT (Delobelle et al., 2020) slightly ahead of the clinical models.

4.2 Unsupervised Representation Analysis

We evaluate the geometric organisation of span representations in the held-out MultiClinNER development set. For each gold entity span, we mean-pool the token hidden states at the span's position to produce a contextual span embedding, leaving 1,350 spans across 120 SNOMED concepts. The results are reported in Table 3.

Retrieval. For each span, we identify its nearest neighbour in the embedding space by cosine similarity, and record whether the neighbour shares the same SNOMED concept. DOKTERBERT retrieves concept-matched neighbours more often than any baseline, while MedRoBERTa.nl-SapBERT's nearest neighbours share a concept less than a third of the time, below both non-grounded models. Same-concept spans end up adjacent in DOKTERBERT's space; in MedRoBERTa.nl-SapBERT's space, applied to contextual spans rather than isolated terms, they do not.

Clustering. We run k-means with $k = 50$ on the span embeddings, then measure how well the resulting clusters align with SNOMED concept labels. Normalised mutual information (NMI) quantifies how much the cluster assignments and the concept labels share, and adjusted Rand index (ARI) measures pairwise agreement corrected for chance. DOKTERBERT's clusters align most closely with SNOMED structure on both metrics. The ARI gap between DOKTERBERT and the rest is larger than the NMI gap, reflecting that the other models produce clusters that are correlated with concept labels but not consistently aligned at the pairwise level.

Entity linking. For each span we compute cosine similarity to its correct SNOMED concept anchor, obtained by encoding the preferred term in isolation, and to a random concept anchor. The discrimi-

	F1			
Model	DISEASE	PROCEDURE	SYMPTOM	Macro
RobBERT	0.7115	0.7410	0.6622	0.7049
MedRoBERTa.nl	0.7116	0.7334	0.6532	0.6994
MedRoBERTa.nl-SapBERT	0.7047	0.7270	0.6493	0.6936
DOKTERBERT	0.7161	0.7310	0.6513	0.6995

Table 2: Supervised NER on MultiClinNER (dev set). Macro F1 averages across the three entity types.

	Retrieval	Clustering		Entity linking	Intra/inter
Model	NN same-concept	NMI	ARI	Discrim. gap	Ratio
RobBERT	0.567	0.835	0.507	+0.045	1.13
MedRoBERTa.nl	0.563	0.824	0.537	+0.108	1.62
MedRoBERTa.nl-SapBERT	0.288	0.715	0.389	+0.170	1.43
DOKTERBERT	**0.607**	**0.903**	**0.729**	**+0.592**	**3.46**

Table 3: Representational analysis on MultiClinNER dev (120 SNOMED concepts, 1,350 contextual spans).

nation gap is the difference. DOKTERBERT's gap exceeds half a cosine similarity unit; every baseline stays below 0.2.

Intra/inter-concept separation. For concepts with at least three spans in the evaluation set, we compute the average pairwise similarity within the same concept and between different concepts. Their ratio captures cluster compactness and separation together. DOKTERBERT's ratio exceeds 3, meaning same-concept spans are more than three times as similar to each other as they are to different-concept spans. The baselines cluster between 1.1 and 1.6, indicating only mild concept-level separation.

5 Discussion

The evaluations give a consistent picture. Supervised fine-tuning overrides pre-training geometry to the point where clinical specialization provides no measurable advantage; further gains on supervised NER will come from scaling the training set or the model, with large language models a plausible next step, not from pre-training innovations at this scale. On the representation evaluations, DOKTERBERT separates from every baseline. MedRoBERTa.nl and RobBERT have no explicit link between their geometry and clinical concept structure; their embedding spaces reflect linguistic co-occurrence rather than clinical meaning. MedRoBERTa.nl-SapBERT trails DOKTERBERT most sharply on the task it was designed for, because its objective aligns isolated terms rather than contextual spans. A representation that strips context to concept identity reduces to symbolic concept lookup; the value of a learned embedding is the contextual information it carries, so that two instances of the same concept qualified differently receive different representations.

The two regimes are complementary. NER extracts clinically informative spans from noisy text; unsupervised representation tasks operate on those spans. Where labelled data is abundant, the starting geometry matters little. In the more common clinical setting where annotation is expensive, inconsistent across sites, or unavailable for rare presentations and novel syndromes, supervised fine-tuning is not a viable option, and representation quality determines what downstream systems can do. Where tasks depend on the embedding space itself, such as similarity-based retrieval, clustering, and anomaly detection over clinical text, the gap is large and observable. Unsupervised representation-based systems for clinical surveillance (Homburg et al., 2026) operate in this regime. DOKTERBERT is the first Dutch clinical language model to align contextual span representations directly to SNOMED concept anchors, providing a representation substrate for clinical applications that cannot rely on task-specific fine-tuning. More broadly, our results show that supervised fine-tuning evaluation obscures differences in representation quality that representation analysis exposes.

Limitations

The contrastive training signal depends on the quality of the span-to-concept linker. Our two-stage linker (exact match against the Dutch SNOMED term set, with a SapBERT similarity fallback at cosine 0.85) introduces errors at both stages: exact match misses surface variants such as abbreviations, spelling variants, and inflected forms, excluding them from training, while the similarity fallback can assign wrong concepts when a near neighbour sits above threshold. We did not evaluate linker precision or recall directly, so the extent of these errors is unknown. Concepts the linker systematically mislinks will have distorted geometry, and concepts it misses entirely will have none. The geometry reflects the linker's decisions, not SNOMED structure in any pure sense.

Our representation evaluations use SNOMED concept identity as ground truth, the same structure used in the contrastive objective. This is appropriate for measuring whether the model encodes clinical concept structure, but does not address whether the geometry transfers to tasks whose label structure differs from SNOMED. The supervised NER evaluation provides one independent check, with DOKTERBERT at parity with baselines on a SNOMED-free task, but extrinsic evaluation across diverse clinical applications is left to future work. A related question concerns coverage: the contrastive objective sees 30,408 SNOMED concepts in training, a small fraction of the full ontology, and whether the geometry generalizes to concepts absent from training is not something our evaluation addresses.

NER results are from single fine-tuning runs without seed averaging. Differences within roughly one F1 point are within typical seed variance; the narrow-band observation is robust to seed choice, but specific orderings within the band are not.

Ethics Statement

This study complies with the Declaration of Helsinki. Data were obtained in a pseudonymised format, with no access to personally identifiable information. The Medical Ethics Review Committee of the University Medical Center Groningen reviewed the study protocol and confirmed that formal ethical approval was not required under Dutch law (reference: METc 2025/357).

Code and Data Availability

The DOKTERBERT pre-trained model is available at `huggingface.co/gijsdanoe/DOKTERBERT` and the training code at `github.com/gijsdanoe/DOKTERBERT`.

The MultiClinNER-nl benchmark is available to shared task participants through the organisers. The pre-training corpus combines openly available Dutch clinical and health-information sources, NTVG content under a formal access agreement with the publisher, and GP consultation notes from the AHON registry. All sources were obtained with formal approval. AHON contains confidential patient health information and cannot be made publicly available; access requires application to the AHON steering committee.

Acknowledgements

The authors gratefully acknowledge Gerlijn Nolle for her role as data manager at AHON (UMCG). We further thank the Nederlands Tijdschrift voor Geneeskunde, the Nederlands Huisartsen Genootschap, Apotheek.nl, the Farmacotherapeutisch Kompas, and the Rijksinstituut voor Volksgezondheid en Milieu for granting permission to use their text in this work.

References

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3255–3265.

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

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Fons Hartendorp, Tom Seinen, Erik van Mulligen, and Suzan Verberne. 2024. Biomedical entity linking for dutch: Fine-tuning a self-alignment bert model on an automatically generated wikipedia corpus. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 253–263.

Maarten Homburg, Gijs Danoe, Marjolein Y. Berger, Tim olde Hartman, Jean Muris, Andreas Voss, Axel Hamprecht, Maarten F. Brilman, Lilian L. Peters, and Matthijs S. Berends. 2026. Ai-driven early infectious disease detection in dutch primary care using bert and ernie. *npj Digital Medicine*, 9:92.

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.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Robin Twickler, Marjolein Y Berger, Feikje Groenhof, Karina Sulim, Liesbeth Ab, Marco H Blanker, Michiel R de Boer, Nynke T Schouwenaars, Guus CGH Blok, and Lilian L Peters. 2024. Data resource profile: Registry of electronic health records of general practices in the north of the netherlands (ahon). *International Journal of Epidemiology*, 53(2):dyae021.

Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the Netherlands*, volume 11, pages 141–159. Computational Linguistics in the Netherlands.

A Hyperparameters

Table 4 reports the full set of training hyperparameters for DOKTERBERT continued pre-training.

Hyperparameter	Value
Base model	MedRoBERTa.nl
Epochs	1
MLM batch size	128
Span batch size	16
Max sequence length	128
Optimizer	AdamW
Learning rate	2×10^{-5}
Warmup steps	1,000
Weight decay	0.01
Gradient clipping	1.0
Contrastive weight α	0.2
Temperature τ	0.07
Graph-distance decay σ	15
Linker similarity threshold	0.85

Table 4: DOKTERBERT continued-pre-training hyperparameters.