DT4H.nl at #SMM4H-HeaRD 2026: Multilingual Clinical NER with multilingual and monolingual models

Bram van Es and DT4H team
University Medical Center Utrecht / Heidelberglaan 100, Utrecht
bes3@umcutrecht

Abstract

We describe the setup we used to complete the MultiClinAI-NER task in the SMM4H-HeaRD workshop 2026. In this work we employed a dedicated multilingual encoder model (EuroBERT-610m), two Dutch encoder models trained from scratch on clinical corpora (MedRoBERTa.nl and CardioDeBERTa.nl) and a generic Dutch encoder model (RobBERT2023-large), all finetuned with a 3-layer DNN head. We find that the use of multilingual datasets is potentially beneficial in augmenting the training corpora of mono-lingual models.

1 Introduction

Clinical named entity recognition (NER) is a core task in biomedical natural language processing (NLP), enabling the extraction of structured information from unstructured clinical narratives. Accurate identification of entities such as diseases, symptoms and medical procedures supports downstream applications including clinical decision support, patient stratification, pharmacovigilance and epidemiological research. While substantial progress has been made for English clinical NLP, many European languages still lack sufficiently large annotated corpora and domain-specific language models to develop robust clinical NER systems.

The MultiClinAI-NER shared task (Gallego-Donoso et al., 2026), organized as part of the SMM4H-HeaRD 2026 workshop (Lopez-Garcia et al., 2026; Gallego-Donoso et al., 2026), addresses this limitation by providing a multilingual benchmark for clinical NER across seven European languages: Czech, Dutch, English, Italian, Romanian, Spanish and Swedish. The dataset combines several parallel clinical corpora annotated for three entity types: *DISEASE*, *SYMPTOM* and *PROCEDURE*. This setting enables the evaluation of both multilingual and monolingual approaches under comparable conditions.

Recent years have seen rapid progress in multilingual transformer architectures. Generic multilingual encoder models such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), EuroBERT (Boizard et al., 2025) and EuroLLM (Martins et al., 2025) demonstrated strong multi-lingual transfer capabilities, while biomedical language models such as BioMistral (Labrak et al., 2024), MMed-Llama (Qiu et al., 2024) and Apollo (Wang et al., 2024) further improved domain adaptation for medical NLP tasks. Cross-lingual transfer learning (Gaschi et al., 2023), trans-tokenization (Remy et al., 2024) and weakly supervised multilingual approaches (Sallauka et al., 2025) have shown that knowledge learned from high-resource languages can benefit lower-resource settings. Nevertheless, it remains unclear to what extent multilingual corpora can strengthen monolingual clinical encoder models, especially when the tokenizer and pretraining corpus are language-specific.

In this work we investigate both multilingual and monolingual approaches for clinical NER. We evaluate the multilingual EuroBERT-610m model alongside several Dutch encoder models, including RobBERT-2023-large (Delobelle and Remy, 2024), MedRoBERTa.nl (Verkijk and Vossen, 2021) and CardioDeBERTa.nl (Van Es, 2026). In particular, we explore whether training a monolingual Dutch clinical model on multilingual annotated corpora can improve generalization across languages despite its language-specific tokenizer and pretraining setup.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 82–87

language	SpaCCC	OnaCCC	CardioCCC
Dutch		1132	508
Spanish	1000		508
Italian		317	508
Swedish		100	508
Romanian		113	508
Czech		100	508
English		1132	508

Table 1: Document distribution

category	Symptom	Disease	Procedure
Dutch	27675	25733	27445
Spanish	29074	26296	28137
Italian	27928	26159	27394
Swedish	27531	25580	27079
Romanian	27015	25561	27313
Czech	27806	25793	27501
English	27465	25118	26733

Table 2: Span counts

Our main contributions are:

- We present a comparative evaluation of multilingual and monolingual transformer-based clinical NER systems on the MultiClinAI dataset.
- We show that multilingual training data leads to a performance of a monolingual clinical encoder model, comparable to a (larger) multilingual model across several languages.

The remainder of this paper describes the dataset, model architectures, training procedure and evaluation results obtained during the shared task.

2 Dataset

The MultiClinAI dataset[1] (Lima López et al., 2026) consists of 7 languages, and is composed of 3 distinct datasets, SpaCCC, OnaCCC and CardioCCC. The dataset contains roughly 25.000 for the categories *symptom*, *disease* and *procedure*, see tables 1 and 2. As of yet, less than half of the documents can be shared across the categories due to slight misalignments.

3 Approach

The code for our model training and inference can be found on Github[2].

[1] https://zenodo.org/records/18508039

[2] https://github.com/DataTools4Heart/CardioNER

Architecture: As pre-trained models we use RobBERT-2023-large (Delobelle and Remy, 2024), CardioDeBERTa.nl (Van Es, 2026) and EuroBERT-610m (Boizard et al., 2025), all BERT-based transformer models. To facilitate an expressive classification layer we replace the single dense linear layer by a 3-layer dense neural network.

Loss: for the multilabel-multiclass model we used a weighted Binary Cross Entropy loss, and for the multiclass model we use a weighted Cross Entropy loss.

Class weighting: to alleviate the impact of a disproportionate amount of O-labels compared to B- and I- labels we use class weights inversely proportional with the relative class count.

Chunking: we applied span-centered chunking, similar to (Van Es et al., 2023). Here, we center the context window around each span and strip to whole sentences. The benefit of this approach is a full exploitation of the available data, a downside is that the training time increases sharply; the number of training windows is basically equal to the number of labeled spans, which could be orders of magnitude larger than the number of paragraph chunks. We set the chunksize and maximum context length to 128 tokens as a trade-off between context length and training time.

Training: We use the BIO-tagging schema. For the training we use a 95% / 5% stratified split. As a matter of experiment we train the Dutch CardioDeBERTa.nl on all languages simultaneously. The idea is that, insofar the tokenization is possible, the model picks up useful information regarding syntax regardless of the language. The use of multiple languages can be seen as a form of data augmentation.
We used a batch size of 16 samples, a learning rate of $2e^{-5}$, and a linear learning rate decay with rate $1e^{-4}$, with 20 epochs of training.

Inference: We use a chunked, token-classification NER pipeline with regex-based pre-tokenization, word-level decoding, optional recovery from uncertain O predictions, entity aggregation based on label certainty and basic ordering logic, followed by a post-hoc span cleanup.

Scoring: We use the F1 score in two scoring regimes, *strict*, and *relaxed*. For strict we require **exact** span overlap and for relaxed scoring **any** overlap counts as a positive.

4 Results

Overall, the competition results can be summarized as: the multilingual EuroBERT performs similarly for all languages ($0.55 - 0.66$ F1 strict). For Dutch the RobBERT2023 model performed slightly better ($0.57 - 0.66$ F1 strict). The Dutch base model CardioDeBERTa was trained using examples from all languages and even outperformed EuroBERT on some languages (Swedish and English) even though it has 200 million fewer parameters and it uses a tokenizer trained on Dutch medical texts. The multilabel multiclass models performed slightly worse than their monolabel multiclass counterparts, most likely because this requires more model expressivity and simultaneously we have less than half the amount of labeled spans available. For all models the relaxed scoring showed scores roughly 0.1 higher.

We note that the chunking approach, the inference strategy, and the model capacity (raw trainable parameters) are determining factors for this work; Specifically, paragraph-based chunking has a lower effective use of context, where part of the corpus is exploited only once. A greedy inference strategy (e.g. the standard approach in the transformers[3] library) might have basic error modes (e.g. mixed B-/I-tags or spuriously missing I-tags) that can be mitigated with custom a-posteriori fixes. A full-weight training implies stronger performance with an increasing number of weights (given enough examples).

We tried the MedRoBERTa.nl (Verkijk and Vossen, 2021) model as well, but it was slightly, yet consistently, underperforming compared to RobBERT2023 and CardioDeBERTa, which we, all other things being equal, associate with the smaller model size. Despite its small size MedRoBERTa.nl performed better than EuroBERT for all Dutch categories on the competition validation set.

For more detailed results we refer to tables 3 to 10.

5 Conclusion

Multilingual NER-models can perform better than strong mono-lingual NER models, and multilingual corpora can be used to train monolingual models. More research is needed to determine to what extent multilingual corpora can be used to strengthen otherwise monolingual models.

6 Caveats

In the inference we accepted only the labeled spans where the scores of **all** the individual B-/I- predictions exceeded a threshold, this likely lowered our recall. Also, the required compute was substantial, limiting the experimental breadth, preventing the extraction of multi-fold statistics, or performing ablation testing, tests with different classification heads or thorough comparisons with multilabel models. Especially, given the short amount of time available. We note that we did not perform model ensembling, or hyperoptimization.

Acknowledgments

The work received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

References

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, and 1 others. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Pieter Delobelle and François Remy. 2024. Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 13:193–203.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[3] https://github.com/huggingface/transformers

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.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical NER: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.

Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. Multiclinai shared task training data.

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.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *Preprint*, arXiv:2408.04303.

Rigon Sallauka, Umut Arioz, Matej Rojc, and Izidor Mlakar. 2025. Weakly-supervised multilingual medical ner for symptom extraction for low-resource languages. *Applied Sciences*, 15(10):5585.

Bram Van Es. 2026. Cardiodeberta.nl$_{clinical}$($revision6bb3528$).

Bram Van Es, Leon C Reteig, Sander C Tan, Marijn Schraagen, Myrthe M Hemker, Sebastiaan RS Arends, Miguel AR Rios, and Saskia Haitjema. 2023. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, 24(1):10.

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.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

A Detailed results, multilingual, strict matching

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.70	0.70	0.71
	EuroBERT-610m	**0.76**	**0.74**	**0.78**
Dutch	CardioDeBERTa	**0.75**	**0.77**	0.72
	EuroBERT-610m	0.71	0.68	**0.75**
English	CardioDeBERTa	0.77	**0.78**	0.77
	EuroBERT-610m	0.77	0.74	**0.80**
Italian	CardioDeBERTa	0.70	0.67	0.72
	EuroBERT-610m	**0.73**	**0.70**	**0.75**
Romanian	CardioDeBERTa	**0.74**	**0.75**	0.72
	EuroBERT-610m	0.71	0.69	**0.73**
Spanish	CardioDeBERTa	0.75	0.73	0.78
	EuroBERT-610m	**0.76**	0.73	**0.80**
Swedish	CardioDeBERTa	0.67	**0.66**	0.68
	EuroBERT-610m	**0.69**	0.65	**0.74**

Table 3: End-to-end span *multilingual* model performance on internal hold-out set with *strict* matching for the DISEASE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.70	0.68	0.72
	EuroBERT-610m	**0.72**	**0.69**	**0.75**
Dutch	CardioDeBERTa	**0.75**	**0.77**	0.74
	EuroBERT-610m	0.73	0.68	**0.78**
English	CardioDeBERTa	**0.81**	**0.80**	0.81
	EuroBERT-610m	0.80	0.77	**0.83**
Italian	CardioDeBERTa	0.62	0.60	0.64
	EuroBERT-610m	**0.65**	**0.62**	**0.68**
Romanian	CardioDeBERTa	0.78	**0.78**	0.77
	EuroBERT-610m	0.78	0.75	**0.82**
Spanish	CardioDeBERTa	0.77	0.74	0.80
	EuroBERT-610m	**0.80**	**0.75**	**0.86**
Swedish	CardioDeBERTa	0.75	**0.75**	0.74
	EuroBERT-610m	**0.76**	0.72	**0.80**

Table 4: End-to-end span *multilingual* performance on internal hold-out set with *strict* matching for the PROCEDURE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.65	0.66	0.65
	EuroBERT-610m	**0.74**	**0.71**	**0.77**
Dutch	CardioDeBERTa	0.71	**0.74**	0.68
	EuroBERT-610m	0.71	0.67	**0.76**
English	CardioDeBERTa	0.79	0.80	0.78
	EuroBERT-610m	**0.81**	0.80	**0.83**
Italian	CardioDeBERTa	0.67	0.67	0.66
	EuroBERT-610m	**0.72**	**0.70**	**0.73**
Romanian	CardioDeBERTa	0.66	**0.70**	0.63
	EuroBERT-610m	**0.69**	0.68	**0.70**
Spanish	CardioDeBERTa	0.73	0.72	0.75
	EuroBERT-610m	**0.78**	**0.74**	**0.82**
Swedish	CardioDeBERTa	0.69	0.71	0.68
	EuroBERT-610m	**0.75**	**0.72**	**0.77**

Table 5: End-to-end span *multilingual* performance on internal hold-out set with *strict* matching for the SYMPTOM category

B Detailed results, multilingual, relaxed matching

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.87	0.86	0.88
	EuroBERT-610m	**0.89** (**0.89**)	0.86	0.91
Dutch	CardioDeBERTa	**0.89**	**0.92**	0.86
	EuroBERT-610m	0.87	0.83	**0.91**
English	CardioDeBERTa	0.89	**0.89**	0.89
	EuroBERT-610m	0.89	0.86	**0.92**
Italian	CardioDeBERTa	0.87	0.84	0.90
	EuroBERT-610m	**0.89**	**0.86**	**0.92**
Romanian	CardioDeBERTa	**0.89**	**0.91**	0.87
	EuroBERT-610m	0.88	0.85	**0.91**
Spanish	CardioDeBERTa	0.87	0.84	0.90
	EuroBERT-610m	**0.88**	0.84	**0.93**
Swedish	CardioDeBERTa	**0.85**	**0.84**	0.86
	EuroBERT-610m	0.84	0.79	**0.90**

Table 6: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the DISEASE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.86	0.83	0.88
	EuroBERT-610m	**0.87**	0.83	**0.91**
Dutch	CardioDeBERTa	**0.90**	**0.92**	0.88
	EuroBERT-610m	0.86	0.81	**0.92**
English	CardioDeBERTa	0.90	**0.90**	0.90
	EuroBERT-610m	0.90	0.87	**0.93**
Italian	CardioDeBERTa	0.87	0.85	0.90
	EuroBERT-610m	**0.89**	0.85	**0.93**
Romanian	CardioDeBERTa	**0.91**	**0.91**	0.90
	EuroBERT-610m	0.90	0.86	**0.94**
Spanish	CardioDeBERTa	0.88	0.84	0.91
	EuroBERT-610m	**0.89**	0.84	**0.95**
Swedish	CardioDeBERTa	**0.89**	**0.89**	0.88
	EuroBERT-610m	0.88	0.84	**0.93**

Table 7: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the PROCEDURE category

language	Base model	F1	Recall	Precision
Czech	CardioDeBERTa	0.83	0.83	0.82
	EuroBERT-610m	**0.87**	**0.84**	**0.91**
Dutch	CardioDeBERTa	**0.86**	**0.89**	0.82
	EuroBERT-610m	0.84	0.79	**0.90**
English	CardioDeBERTa	0.89	**0.90**	0.88
	EuroBERT-610m	0.89	0.88	**0.91**
Italian	CardioDeBERTa	0.84	0.85	0.84
	EuroBERT-610m	**0.88**	**0.86**	**0.90**
Romanian	CardioDeBERTa	0.81	**0.86**	0.77
	EuroBERT-610m	**0.83**	0.82	**0.85**
Spanish	CardioDeBERTa	0.84	0.82	0.86
	EuroBERT-610m	**0.86**	0.82	**0.91**
Swedish	CardioDeBERTa	0.85	**0.87**	0.83
	EuroBERT-610m	**0.87**	0.84	**0.90**

Table 8: End-to-end span *multilingual* performance on internal hold-out set with *relaxed* matching for the SYMPTOM category

C Detailed results, Dutch

category	language	Base model	F1	Recall	Precision
DISEASE	multilingual	CardioDeBERTa	**0.75**	**0.77**	0.72
		EuroBERT-610m	0.71	0.68	**0.75**
	monolingual	RobBERT2023-large	0.67	0.68	0.67
		MedRoBERTa.nl	0.64	0.68	0.61
PROCEDURE	multilingual	CardioDeBERTa	**0.75**	**0.77**	0.74
		EuroBERT-610m	0.73	0.68	**0.78**
	monolingual	RobBERT2023-large	0.71	0.73	0.7
		MedRoBERTa.nl	0.65	0.68	0.62
SYMPTOM	multilingual	CardioDeBERTa	0.71	**0.74**	0.68
		EuroBERT-610m	0.71	0.67	**0.76**
	monolingual	RobBERT2023-large	0.63	0.64	0.62
		MedRoBERTa.nl	0.61	0.65	0.56

Table 9: End-to-end span performance on internal hold-out set with *strict* matching

category	language	Base model	F1	Recall	Precision
DISEASE	multilingual	CardioDeBERTa	**0.89**	**0.92**	0.86
		EuroBERT-610m	0.87	0.83	**0.91**
	monolingual	RobBERT2023-large	0.85	0.86	0.85
		MedRoBERTa.nl	0.86	0.9	0.82
PROCEDURE	multilingual	CardioDeBERTa	**0.90**	**0.92**	0.88
		EuroBERT-610m	0.86	0.81	**0.92**
	monolingual	RobBERT2023-large	0.86	0.88	0.85
		MedRoBERTa.nl	0.85	0.89	0.81
SYMPTOM	multilingual	CardioDeBERTa	**0.86**	**0.89**	0.82
		EuroBERT-610m	0.84	0.79	**0.90**
	monolingual	RobBERT2023-large	0.81	0.83	0.79
		MedRoBERTa.nl	0.8	0.87	0.75

Table 10: End-to-end span performance on internal hold-out set with *relaxed* matching