SIEMENS at #SMM4H–HeaRD 2026: The Impact of Training Strategy and Backbone Selection on BERT-based Multilingual Clinical NER

Manuela Daniela Danu[1, 2],

[1]Foundational Technologies, Siemens AG, Brașov, Romania,
[2]Automation and Information Technology, Transilvania University of Brașov, Brașov, Romania,
manuela.voinea@siemens.com

Abstract

This paper describes our participation in the MultiClinNER subtask of the MultiClinAI shared task, part of the #SMM4H-HeaRD Workshop at ACL 2026. The task requires identifying DISEASE, SYMPTOM, and PROCEDURE mentions in clinical case reports across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. We compare two BERT-based sequence labeling methods: (i) sentence-level token classification with a fixed train/validation split, and (ii) paragraph-level chunking with 5-fold cross-validation and checkpoint merging, using language-specific BERT models and multilingual XLM-RoBERTa-large as backbones. Our results show that 5-fold training with checkpoint merging consistently outperforms the fixed split strategy, with further analysis suggesting that the gains are primarily driven by improved training-set coverage rather than by differences in input granularity. Language-specific BERT encoders prove most effective for Spanish and English, while XLM-RoBERTa-large yields the strongest results for the remaining five languages through cross-lingual transfer.

1 Introduction

The extraction of structured information from free-text clinical narratives is a long-standing goal of clinical natural language processing (NLP), enabling downstream applications such as clinical decision support (Eguia et al., 2024), disease surveillance (Methuku, 2025), cohort identification, and patient phenotyping from electronic health records (EHRs) (Shivade et al., 2014). Named entity recognition (NER) is a core task in clinical information extraction, which involves identifying and classifying mentions of clinical concepts in text, such as diseases, medications, symptoms, procedures, and laboratory results (Wu et al., 2018).

Recent advances in pre-trained language models have substantially improved clinical NER, particularly in English (Lee et al., 2020; Alsentzer et al., 2019). However, extending these advances to other languages remains challenging because high-quality, expert-annotated clinical corpora are expensive to create and require substantial domain expertise. While such resources are already limited in monolingual settings, especially for low- and mid-resource languages (Névéol et al., 2018; Shaitarova et al., 2023), multilingual corpora annotated under consistent guidelines are even scarcer (Rodríguez-Miret et al., 2024), constraining the development of robust multilingual clinical NER systems and, more broadly, limiting cross-lingual clinical research and data analysis.

To address this gap, recent work has explored cross-lingual transfer strategies based on multilingual pre-trained models (Conneau et al., 2020), machine translation (Gaschi et al., 2023), and annotation projection (Agerri et al., 2018). These approaches provide a practical path toward multilingual clinical NER by transferring supervision from resource-rich to resource-scarce languages, for example by translating annotated source-language corpora, projecting entity labels into target languages, and validating them with domain experts (Miranda-Escalada et al., 2022; Lima-López et al., 2023). However, their effectiveness for fine-grained clinical NER across typologically diverse languages has not yet been comprehensively benchmarked.

The MultiClinAI shared task (Gallego-Donoso et al., 2026), organized as part of the 11th Social Media Mining for Health Applications and Health Real-World Data (#SMM4H-HeaRD) 2026 Workshop (Lopez-Garcia et al., 2026) at the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), provides a rigorous benchmark for multilingual clinical NER. Its first subtask, MultiClinNER, focuses on identifying DISEASE, SYMPTOM, and PROCEDURE mentions in clinical case reports across seven lan-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 216–221

guages from the Romance, Germanic, and Slavic families: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. The dataset draws on three clinical case report corpora (SpaCCC, CardioCCC, and OnaCCC) (Lima López et al., 2026), and includes both machine-translated texts with projected, expert-validated annotations and documents originally written in the target languages. This combination of translated and native texts provides a realistic evaluation setting for multilingual clinical NER.

In this paper, we describe our participation in the MultiClinNER subtask. We formulate NER as a sequence labeling task using BIO tags (Ramshaw and Marcus, 1999) and fine-tune two families of transformer encoders for token classification: language-specific BERT models (Devlin et al., 2019) pre-trained on cardiology-domain text and XLM-RoBERTa (Conneau et al., 2020). The BERT models are fine-tuned separately for each language, while XLM-RoBERTa is fine-tuned in both monolingual and multilingual settings. This setup allows us to assess the impact of domain pre-training and the potential benefits of multilingual training on clinical NER.

2 Related Work

Recent shared tasks have progressively extended clinical NER beyond English. The DisTEMIST shared task (Miranda-Escalada et al., 2022) introduced a gold-standard corpus of 1,000 Spanish clinical case reports annotated for disease mentions and also released multilingual silver-standard versions in seven languages, generated through neural machine translation and annotation projection. MedProcNER (Lima-López et al., 2023) extended this line of work to clinical procedures by providing a gold-standard Spanish corpus on the same document collection, alongside silver-standard resources for eight languages. Building on these efforts, MultiCardioNER (Lima-López et al., 2024) was the first shared task to evaluate multilingual clinical medication recognition in cardiology case reports across Spanish, English, and Italian, using translated corpora whose projected annotations were manually validated by bilingual clinical experts. MultiClinAI (Gallego-Donoso et al., 2026) further expands this paradigm to seven languages and three entity types, yielding the broadest multilingual clinical NER benchmark currently available.

3 Methods

3.1 Datasets

The MultiClinNER corpus (Lima López et al., 2026) consists of clinical case reports in seven languages: Czech (cz), Dutch (nl), English (en), Italian (it), Romanian (ro), Spanish (es), and Swedish (sv). It is annotated for three entity types – DISEASE, SYMPTOM, and PROCEDURE – with entity spans and character offsets encoded in BRAT format.

The corpus builds on three sources. SpaCCC (Spanish Clinical Case Corpus), which contains 1,000 multi-specialty clinical case reports, and CardioCCC (Cardiology Clinical Case Corpus), which contains 508 cardiology case reports, were both originally annotated in Spanish and later extended to six target languages through machine translation, lexical annotation projection, and validation by bilingual clinical experts. OnaCCC (Original Native Clinical Case Corpus) complements these translated resources with reports originally written in each target language, sourced from open-access journals and annotated following the same guidelines. Together, the translated and native texts provide a comprehensive multilingual evaluation setting.

3.2 Experiments

We formulate NER as a BIO sequence labeling task and explore two approaches, each evaluated in monolingual and multilingual settings. Method 1 serves as a straightforward baseline, while Method 2 is motivated by two hypotheses: (1) broader input context may help disambiguate clinical entities, and (2) exposing the model to the full training set through checkpoint merging may be especially beneficial in resource-constrained settings.

3.2.1 Method 1: Sentence-level token classification

This approach converts BRAT annotations to CoNLL-format, fine-tunes a BERT-based encoder with a linear token-classification head, processing one sentence at a time, and maps predicted BIO labels back to BRAT spans at inference time, as described in (Danu et al., 2025). In the monolingual setting, separate models are trained per entity type and per language (21 models in total), using CardioBERTa (DataTools4Heart, 2025) – a family of seven language-specific encoders pretrained on general, biomedical, and cardiology

text – as base models. In the multilingual setting, XLM-RoBERTa-large (Conneau et al., 2020) is used as a shared backbone trained on concatenated multilingual data, resulting in 3 multilingual models, one for each entity type. The training data was split into fixed train and validation sets (80% and 20%, respectively). Consequently, models were exposed to only a fraction of the available annotations during training.

3.2.2 Method 2: Paragraph-level sequence labeling

This pipeline operates on document-level annotations and chunks long clinical reports into paragraph-level segments before encoding. Models are trained per entity type using 5-fold cross-validation, in both multiclass (softmax over BIO labels) and multilabel (independent sigmoid per label) configurations. The resulting fold checkpoints are averaged into a single merged model, ensuring that the final model has effectively been trained on the entire training set despite individual folds seeing only a subset. In the monolingual setting, models are trained with both CardioBERTa and XLM-RoBERTa-large backbones, while in the multilingual setting, only XLM-RoBERTa-large is used.

3.2.3 Training details

All models were trained on an NVIDIA RTX A5000 GPU (24 GB) for 10 epochs with a maximum sequence length of 256 tokens. Method 1 used a batch size of 8, a learning rate of 9×10^{-6}, and a weight decay of 10^{-2}, whereas Method 2 used a batch size of 16, a learning rate of 7×10^{-5}, and a weight decay of 10^{-4}. Performance was measured using the official MultiClinAI evaluation scripts, which report precision, recall, and F1-score under two settings: strict matching, where a prediction is correct only if the entity label and character offsets exactly match the gold annotation, and character-overlap, which measures span agreement at the character level.

4 Results

Table 1 reports the best strict and character-overlap F1 scores achieved with Method 1 and Method 2 on the MultiClinNER test set for each language and entity type. Method 2 consistently outperforms Method 1 across all 21 language–entity pairs, with strict F1 improvements ranging from 1.19 to 5.64 percentage points and averaging 2.51 points. English shows the largest average improvement (+4.12 points), while Dutch (+1.77) and Swedish (+1.90) show the smallest gains.

The two methods differ in two key aspects. First, they rely on different training strategies: Method 2 uses 5-fold cross-validation with checkpoint merging, whereas Method 1 uses a fixed 80/20 train–validation split. Second, they differ in input granularity: Method 2 processes paragraph-level chunks, while Method 1 operates at the sentence level. To disentangle the individual contributions of these two factors, we additionally compare validation-set results. When trained on a comparable data fraction of 80%, Method 1 slightly outperforms individual Method 2 folds by an average of 1.92 strict F1 points across all 21 language–entity cases. Despite non-identical validation splits, which limit the strength of this comparison, the consistency of the performance gap in favour of Method 1 across all 21 language–entity pairs suggests that paragraph-level context does not, by itself, explain the test-set gains of Method 2. The test-set improvements are therefore more plausibly attributable to the fold-wise training and checkpoint-merging strategy, which effectively exposes the final model to the entire training set.

Since all top results in Table 1 were obtained with Method 2 using 5-fold merged models, we restrict the remainder of our analysis to Method 2. Spanish, the source language in which the majority of the corpus was written and annotated, achieves the highest scores across all entity types, with strict F1 ranging from 74.67% to 80.69%, followed by English (70.27% – 78.15%) and Romanian (67.34% – 73.18%). Dutch yields the lowest overall performance, with an average strict F1 of 66.6% across entities, followed by Czech (67.46%), Italian (67.75%), and Swedish (68.52%). Across entity types, DISEASE and PROCEDURE show broadly comparable performance, whereas SYMPTOM is consistently the most challenging one. One possible explanation is that symptoms are expressed with greater linguistic variability, while diseases and procedures tend to be described using more standardized terminology.

With respect to backbone selection, multilingual XLM-RoBERTa-large achieves the best results in 16 of 21 cases, with CardioBERTa outperforming it only in Spanish across all entity types and in English for DISEASE and SYMPTOM entities. This suggests that domain-specific pre-training provides a clear advantage only where substantial

		Method 1			Method 2		
Entity	**Lang**	**Strict F1**	**Char F1**	**Model**	**Strict F1**	**Char F1**	**Model**
DIS	**cz**	0.6639	0.7821	XLM-R-L_Multi	**0.6926**	0.7932	MC_XLM-R-L_Multi
	en	0.7301	0.8275	CardioBERTa	**0.7815**	0.8469	MC_CardioBERTa
	es	0.7865	0.8586	CardioBERTa	**0.8069**	0.8605	MC_CardioBERTa
	it	0.6922	0.8150	CardioBERTa	**0.7086**	0.8034	MC_XLM-R-L_Multi
	nl	0.6831	0.7832	XLM-R-L_Multi	**0.6974**	0.7895	MC_XLM-R-L_Multi
	ro	0.7127	0.8245	XLM-R-L_Multi	**0.7298**	0.8264	ML_XLM-R-L_Multi
	sv	0.6744	0.7825	CardioBERTa	**0.6923**	0.7964	MC_XLM-R-L_Multi
PROC	**cz**	0.6684	0.8078	XLM-R-L_Multi	**0.7007**	0.8121	MC_XLM-R-L_Multi
	en	0.6972	0.8186	CardioBERTa	**0.7288**	0.8258	MC_XLM-R-L_Multi
	es	0.7806	0.8652	CardioBERTa	**0.7925**	0.8554	MC_CardioBERTa
	it	0.6344	0.7898	CardioBERTa	**0.6908**	0.8079	ML_XLM-R-L_Multi
	nl	0.6755	0.7874	XLM-R-L_Multi	**0.6977**	0.8004	MC_XLM-R-L_Multi
	ro	0.6967	0.8184	XLM-R-L_Multi	**0.7318**	0.8324	MC_XLM-R-L_Multi
	sv	0.6876	0.8034	XLM-R-L_Multi	**0.7052**	0.8203	ML_XLM-R-L_Multi
SYMP	**cz**	0.6182	0.7407	XLM-R-L_Multi	**0.6306**	0.7477	MC_XLM-R-L_Multi
	en	0.6621	0.7685	CardioBERTa	**0.7027**	0.7803	MC_CardioBERTa
	es	0.7210	0.8124	CardioBERTa	**0.7467**	0.8106	MC_CardioBERTa
	it	0.6168	0.7607	XLM-R-L_Multi	**0.6332**	0.7694	ML_XLM-R-L_Multi
	nl	0.5863	0.7186	XLM-R-L_Multi	**0.6030**	0.7228	MC_XLM-R-L_Multi
	ro	0.6528	0.7701	XLM-R-L_Multi	**0.6734**	0.7741	MC_XLM-R-L_Multi
	sv	0.6365	0.7613	XLM-R-L_Multi	**0.6581**	0.7742	MC_XLM-R-L_Multi

Table 1: Best results on the MultiClinNER test set for Method 1 and Method 2. For each language and entity type, the model with the highest strict F1 is reported alongside character-level F1 and model name. **Bold** indicates the best strict F1 across both methods. DIS = DISEASE; PROC = PROCEDURE; SYMP = SYMPTOM; XLM-R-L = XLM-RoBERTa-large; MC = multiclass; ML = multilabel; Multi = multilingual.

target-language data was available, while cross-lingual transfer compensates effectively in all other settings. Among model configurations, multiclass outperforms multilabel in 17 of 21 cases, indicating that standard softmax training over BIO labels is generally sufficient for single-entity NER.

The character-overlap F1 score exceeds strict F1 by approximately 10 points on average, with smaller gaps for Spanish and English and larger gaps for languages such as Italian, Swedish, Czech, and Dutch. This indicates that many predictions identify the correct entity type and approximate span location but fail to match the exact span boundaries, suggesting that boundary detection, rather than entity classification, remains the main source of error.

5 Conclusions

We compared two BERT-based sequence labeling methods for the MultiClinNER subtask across seven languages and three entity types. Our results show that 5-fold training with checkpoint merging (Method 2) consistently outperforms a fixed train/validation split strategy (Method 1), underscoring the value of maximizing effective training set coverage for clinical NER. Further analysis suggests that these gains are driven primarily by the fold-wise training and checkpoint-merging strategy, rather than by differences in input granularity.

Domain-specific BERT-based encoders performed best for Spanish and English, where substantial pre-training data was available, whereas multilingual XLM-RoBERTa-large achieved the strongest results for the remaining five languages, highlighting the effectiveness of cross-lingual transfer in lower-resource settings. The persistent gap between character-overlap and strict F1 scores suggests that exact boundary detection, rather than entity classification, remains a major challenge.

Overall, our results demonstrate that competitive multilingual clinical NER is achievable with a relatively simple fine-tuning pipeline when backbone selection is aligned with resource availability and training strategies are designed to maximize the use of available data.

Limitations

Our work has several limitations. First, we restricted our experiments to BERT-based encoder models and did not investigate generative or large language model approaches, which may offer complementary strengths for clinical NER. Second, hyperparameter tuning was limited, since all models were trained with a fixed configuration for each method, and systematic optimization could yield further gains. Third, our comparison between Method 1 and individual Method 2 folds on the validation set relies on non-identical data splits, which limits the strength of our conclusions regarding the relative contribution of input granularity versus training strategy. Fourth, our error analysis primarily focused on the gap between strict and character-overlap F1 scores and did not include a fine-grained qualitative analysis, which could reveal the specific linguistic patterns driving boundary mismatches. Finally, although we evaluated our approach across seven languages, the non-Spanish training data was derived primarily from Spanish through translation and annotation projection, which may introduce translation artifacts that affect model performance unevenly across target languages.

Acknowledgments

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

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

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.

Manuela Daniela Danu, George Marica, Constantin Suciu, Lucian Mihai Itu, and Oladimeji Farri. 2025. Multilingual clinical ner for diseases and medications recognition in cardiology texts using bert embeddings. *arXiv preprint arXiv:2510.17437*.

DataTools4Heart. 2025. CardioBERTa Family. Hugging Face collection. Accessed April 27, 2026.

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.

Hans Eguia, Carlos Luis Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, and Francesc Saigí-Rubió. 2024. Clinical decision support and natural language processing in medicine: systematic literature review. *Journal of Medical Internet Research*, 26:e55315.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In *CLEF (Working Notes)*, pages 1–18.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossoff, Anoop Shah, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2024. Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In *CLEF Working Notes*.

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.

Vijayalaxmi Methuku. 2025. Nlp and ai for public health intelligence: Automating disease surveillance from unstructured data. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(1):43–56.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Jan Rodríguez-Miret, Eulàlia Farré-Maduell, Salvador Lima-López, Laura Vigil, Vicent Briva-Iglesias, and Martin Krallinger. 2024. Exploring the potential of neural machine translation for cross-language clinical natural language processing (nlp) resource generation through annotation projection. *Information*, 15(10):585.

Anastassia Shaitarova, Jamil Zaghir, Alberto Lavelli, Michael Krauthammer, and Fabio Rinaldi. 2023. Exploring the latest highlights in medical natural language processing across multiple languages: a survey. *Yearbook of medical informatics*, 32(01):230–243.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2018. Clinical named entity recognition using deep learning models. In *AMIA annual symposium proceedings*, volume 2017, page 1812.