# SINAI at #SMM4H–HeaRD 2026: Multilingual Clinical NER with MrBERT-biomed and Optuna Hyperparameter Optimization

**Lucas Molino Piñar and Manuel Carlos Díaz Galiano and María Teresa Martín Valdivia**
SINAI Research Group, Universidad de Jaén, Spain
{lmolino, mcdiaz, maite}@ujaen.es

## Abstract

This paper describes the system submitted by our team to the MultiClinAI shared task at the 11th SMM4H-HeaRD Workshop (ACL 2026). The task addresses multilingual clinical Named Entity Recognition (NER) for three entity types (DISEASE, PROCEDURE, and SYMPTOM) in Spanish clinical texts. Our approach fine-tunes **MrBERT-biomed**, a domain-adapted ModernBERT model pre-trained on biomedical corpora, using multilingual clinical data from seven European languages. We train independent entity-specific models, each optimized via Bayesian hyperparameter search with Optuna, and apply a deterministic post-processing step that aligns predicted spans to word boundaries. On the official test set, our system achieves overall strict micro-F1 scores of 0.7453, 0.7107, and 0.6603 for DISEASE, PROCEDURE, and SYMPTOM, respectively.

## 1 Introduction

The automatic extraction of clinical entities from electronic health records (EHRs) is a fundamental task in biomedical natural language processing (NLP). Accurate identification of diseases, symptoms, and medical procedures enables downstream applications such as clinical decision support, pharmacovigilance, and epidemiological surveillance.

The MultiClinAI shared task, organized within the 11th Social Media Mining for Health and Health Real-World Data (SMM4H-HeaRD) Workshop at ACL 2026 (Lopez-Garcia et al., 2026; Lima-López et al., 2026a), proposes a multilingual clinical NER challenge over the MultiClinNER corpus (Lima-López et al., 2026b). Participants are provided with clinical documents annotated in seven European languages (Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish) covering three entity types: DISEASE, PROCEDURE, and SYMPTOM. The evaluation focuses on Spanish, using strict span-level micro-F1 as the primary metric.

In this paper we describe the system submitted by our team. Our approach fine-tunes **MrBERT-biomed** (Tamayo et al., 2026), a recent biomedically adapted variant of the ModernBERT architecture (Warner et al., 2025), on the full multilingual training set with automated hyperparameter optimization via Optuna (Akiba et al., 2019). A rule-based post-processing step corrects span boundaries at the word level to alleviate the impact of subword tokenization on the strict evaluation metric.

## 2 System Description

Figure 1 provides an overview of our system pipeline, from multilingual data ingestion to final span-level prediction.

### 2.1 Pre-trained Language Model

Our system builds upon **MrBERT-biomed**, developed by the Barcelona Supercomputing Center (BSC-LT). MrBERT-biomed is a multilingual encoder based on the ModernBERT architecture (Warner et al., 2025), obtained through continued pre-training of MrBERT-es (a general-purpose multilingual encoder) on approximately 24 billion tokens of biomedical literature. The model supports a context window of up to 8,192 tokens, with pre-training data composed predominantly of English (84.7%) and Spanish (14.8%) biomedical text alongside smaller proportions of other European languages. The choice of MrBERT-biomed over the general-purpose baseline is motivated by our preliminary experiments (Section 3), which showed that biomedical domain adaptation provides a consistent and meaningful advantage for clinical entity recognition.

### 2.2 Task Formulation

We formulate clinical NER as a BIO sequence labeling task (Ramshaw and Marcus, 1995). For each entity type



*Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks*, pages 18–22

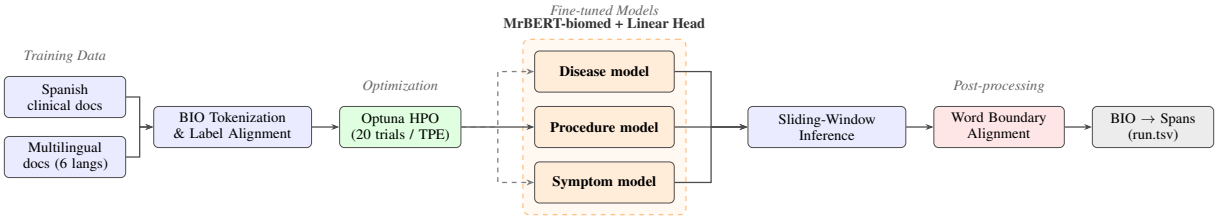


Figure 1: Overview of the proposed system architecture. Multilingual clinical documents are tokenized with BIO label alignment, then three independent entity-specific models (Disease, Procedure, Symptom) are fine-tuned from MrBERT-biomed with Optuna-driven hyperparameter optimization. At inference time, sliding-window decoding and word boundary alignment produce the final span-level predictions.

$e \in \{\text{DISEASE}, \text{PROCEDURE}, \text{SYMPTOM}\}$, the label set is $\mathcal{L}_e = \{\texttt{O}, \texttt{B-}e, \texttt{I-}e\}$, and a linear classification head followed by softmax decoding is placed on top of the MrBERT-biomed encoder.

### 2.3 Training Strategy

**Entity-Specific Models.** Rather than training a single model for all three entity types simultaneously, we train **three independent models**, one per entity type. Each model is initialized from the same MrBERT-biomed checkpoint and fine-tuned exclusively on its corresponding entity annotations. This design avoids potential label interference between entity types and allows each model to receive independently optimized hyperparameters.

**Multilingual Training Data.** During development we evaluated three language strategies: Spanish-only training, full multilingual training over all seven languages, and combinations thereof. Incorporating multilingual data consistently improved performance over the Spanish-only configuration for both model variants (Section 3), likely due to the increased volume and diversity of clinical annotations. Consequently, our final submission uses all available multilingual data for training while restricting validation and test evaluation to Spanish.

**Hyperparameter Optimization with Optuna.** For each entity-specific model we conduct automated hyperparameter optimization (HPO) using Optuna (Akiba et al., 2019). The search space covers learning rate ($[10^{-5}, 5{\times}10^{-4}]$, log-uniform), weight decay ($[0.0, 0.3]$, uniform), warmup ratio ($[0.0, 0.2]$, uniform), batch size ($\{4, 8, 16, 32\}$, categorical), number of epochs ($[3, 6]$, integer), and gradient accumulation steps ($\{1, 2, 4\}$, categorical). Each HPO run consists of 20 trials with a Tree-structured Parzen Estimator (TPE) sampler and a median pruner, maximizing the seqeval micro-F1 score on the Spanish validation set. Table 1 summarizes the best hyperparameters identified for each entity type.

| Hyperparameter | Disease | Procedure | Symptom |
|---|---|---|---|
| Learning rate | 4.9e-4 | 3.4e-4 | 2.4e-4 |
| Epochs | 10 | 7 | 9 |
| Batch size | 8 | 4 | 4 |

Table 1: Optimal hyperparameters identified by Optuna for each entity-specific model using MrBERT-biomed with multilingual training.

**Training Infrastructure.** All models are fine-tuned with the HuggingFace `Transformers` library (Wolf et al., 2020) on a single NVIDIA Ampere GPU (40 GB VRAM) using mixed-precision training (BF16). The maximum sequence length is set to 512 tokens with a stride of 128 for sliding-window inference on longer documents.

### 2.4 Post-Processing: Word Boundary Alignment

Subword tokenization can produce entity spans that do not align with word boundaries, leading to partial-word predictions penalized under the strict evaluation metric. To address this, we apply a deterministic post-processing step at inference time that first strips leading and trailing non-alphanumeric characters from the predicted span, then expands it leftward and rightward to the nearest word boundary, where a word character is defined as any alphanumeric character including Spanish diacritics (*á, é, í, ó, ú, ü, ñ*). This heuristic is computationally inexpensive and operates directly on the original document text.

| Strategy | Disease | Proced. | Symptom |
|---|---|---|---|
| MrBERT-es (Multi.) | 0.7439 | **0.7625** | 0.7232 |
| MrBERT-bio (ES) | 0.7107 | 0.7207 | 0.6863 |
| MrBERT-bio (Multi.) | **0.7478** | 0.7597 | **0.7345** |

Table 2: Development results (strict micro-F1) on the Spanish validation set for different model and data strategies.

| Entity | Corpus | P | R | F1 |
|---|---|---|---|---|
| DISEASE | Overall | 0.746 | 0.744 | **0.745** |
| | Native | 0.697 | 0.734 | 0.715 |
| PROCEDURE | Overall | 0.746 | 0.679 | **0.711** |
| | Native | 0.723 | 0.501 | 0.592 |
| SYMPTOM | Overall | 0.685 | 0.637 | **0.660** |
| | Native | 0.704 | 0.505 | 0.588 |

Table 3: Official test results (strict precision, recall, F1) for our submitted system.

## 3 Experimental Setup and Results

### 3.1 Development Experiments

During development we systematically compared three configurations to assess the relative contribution of domain adaptation and multilingual data augmentation: (i) **MrBERT-es (Multilingual)**, a general-purpose encoder trained on multilingual clinical data; (ii) **MrBERT-biomed (Spanish-only)**, a domain-adapted encoder trained exclusively on Spanish; and (iii) **MrBERT-biomed (Multilingual)**, the domain-adapted encoder trained on all seven languages. Table 2 reports strict micro-F1 scores on the Spanish validation set for each configuration.

Two conclusions emerge from these results. First, multilingual training consistently outperforms the Spanish-only configuration for both model variants, confirming that cross-lingual annotation diversity provides a useful inductive bias even when evaluating only in Spanish. Second, combining biomedical domain adaptation with multilingual data yields the best overall performance across entity types, with MrBERT-biomed (Multilingual) achieving the highest F1 scores for DISEASE and SYMPTOM and a competitive result for PROCEDURE. This configuration was therefore selected as our final submission.

### 3.2 Official Test Results

Table 3 presents our official results on the test set, broken down by entity type and corpus type (Supervised Translated vs. Native).

DISEASE is the best-performing entity type (F1 = 0.745), likely because disease mentions exhibit more regular morphological patterns and benefit from the highest annotation density in the training data. A consistent drop in recall on the Native subcorpus relative to the Supervised Translated subcorpus is observed across all entity types, and is particularly pronounced for PROCEDURE (0.501) and SYMPTOM (0.505). This gap suggests that natively authored clinical documents contain more diverse and challenging entity expressions (idiomatic phrasings, abbreviations, and domain-specific jargon) than their machine-translated counterparts, a pattern we examine further in Section 4.

## 4 Analysis

**Benefit of Biomedical Pre-training.** The approximately 24 billion tokens of biomedical text used during continued pre-training of MrBERT-biomed provide the encoder with a semantic prior over medical terminology that complements task-specific fine-tuning. Our development experiments confirm that this domain knowledge yields consistent improvements over the general-purpose MrBERT-es baseline, particularly for lower-frequency entity types where task-specific data is sparser.

**Multilingual Data as Regularization.** Expanding the training set to seven languages effectively increases corpus diversity and acts as a form of implicit regularization. Despite evaluating exclusively in Spanish, the multilingual signal provides structural and terminological variety that helps reduce overfitting to the patterns of a single language, with the most pronounced benefits appearing for SYMPTOM, where recall improves by almost five absolute points compared to the Spanish-only configuration.

**Error Analysis: Native vs. Translated Corpora.** The most salient pattern in our results is the performance gap between the translated and native subcorpora. Recall on the native subcorpus falls to 0.501 for PROCEDURE and 0.505 for SYMPTOM, indicating that natively authored clinical documents contain idiomatic expressions, abbreviations, and domain-specific jargon that are underrepresented in the training data. This observation motivates future work on domain-specific data augmentation and few-shot adaptation strategies targeted at native clinical writing styles.

**Limitations of Post-Processing.** The word boundary alignment step improves strict F1 by cor-

recting partial-word predictions resulting from subword tokenization, but it cannot recover entities missed entirely by the model. Furthermore, the heuristic is tailored to Spanish orthography and would require adaptation for other target languages, limiting its generalizability in a fully multilingual deployment.

## 5 Conclusion

We have presented our system for the MultiClinAI shared task at SMM4H-HeaRD 2026, addressing multilingual clinical NER in Spanish. Our approach combines three key design choices: (i) leveraging MrBERT-biomed, a domain-adapted ModernBERT encoder pre-trained on 24 billion biomedical tokens; (ii) training independent entity-specific models with Optuna-driven Bayesian hyperparameter optimization; and (iii) applying a deterministic word boundary alignment post-processing step to improve strict span-level evaluation. The system achieves strict micro-F1 scores of 0.745, 0.711, and 0.660 for DISEASE, PROCEDURE, and SYMPTOM, respectively.

Our experiments demonstrate that biomedical domain adaptation and multilingual training data are complementary strategies that together yield the strongest performance across all entity types. The analysis also reveals a significant performance gap between translated and natively authored clinical texts, with recall dropping considerably on the native subcorpus, particularly for PROCEDURE and SYMPTOM.

Future work will explore joint multi-task architectures to exploit inter-entity dependencies, ensemble methods, and domain-specific data augmentation techniques targeting native clinical writing styles to bridge the observed performance gap.

## Limitations

Our system has several limitations worth acknowledging. Training independent models per entity type increases computational cost and does not exploit potential inter-entity dependencies; a joint multi-task architecture could share representations across entity types and may improve efficiency. We also do not employ ensemble methods in our final submission, although preliminary experiments with 5-fold cross-validation showed marginal improvements. The post-processing heuristic is deterministic and language-specific, and does not handle multi-word expressions with embedded punctuation. Finally, the substantial performance gap on the native subcorpus suggests that domain-specific data augmentation or few-shot adaptation strategies could substantially improve robustness on natively authored clinical texts.

## Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project CONSENSO (PID2021-122263OB-C21), Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by Junta de Andalucía.

## References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Salvador Lima-López, Fernando Gallego, and 1 others. 2026a. Overview of the MultiClinAI Shared Task at SMM4H-HeaRD 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.

Salvador Lima-López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026b. 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.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *Computing Research Repository*, arXiv:cmp-lg/9505040.

Daniel Tamayo, Iñaki Lacunza, Paula Rivera-Hidalgo, Severino Da Dalt, Javier Aula-Blasco, Aitor Gonzalez-Agirre, and Marta Villegas. 2026. Mr-bert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation. *Preprint*, arXiv:2602.21379.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.