The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, Martin Krallinger
Barcelona Supercomputing Center,
{fgalleg1,slimalop,efarre,jrosell,mkrallin}@bsc.es

Abstract

We present an overview of the MultiClinAI shared task, which focuses on multilingual clinical entity extraction and automatic corpus generation through annotation projection. It addresses two key challenges in clinical natural language processing (NLP): (i) developing comparable multilingual named entity recognition (NER) systems and (ii) automatically constructing multilingual clinical corpora through annotation projection. The MultiClinAI task provides a unified benchmark for evaluating multilingual and cross-lingual clinical NLP approaches that cover diseases, symptoms, and procedures in Spanish, English, Dutch, Italian, Romanian, Swedish, and Czech. A total of 21 teams from 13 countries participated, submitting 531 runs across the different subtasks. The top runs obtained very competitive results across several languages and entity types. The results highlight both the challenges and opportunities of multilingual clinical information extraction. All resources, including a corpus of over 738,201 manually revised entity mentions across seven languages, are publicly available on Zenodo at: https://zenodo.org/records/19334278.

1 Introduction

Named entity recognition (NER) systems play a key role in clinical natural language processing (NLP) applications by identifying essential clinical variables or concept types—such as diseases and comorbidities, medications, signs and symptoms or clinical procedures—in medical documents and EHRs. Such systems have proven valuable in optimizing clinical workflows, enhancing decision support, and supporting large-scale health data analysis (Wang et al., 2018). However, the development and evaluation of robust clinical NER systems depends heavily on the availability of well-annotated corpora. These corpora must be carefully curated by clinical experts to ensure accuracy—a process that is both time-consuming and expensive. Moreover, annotated corpora are often language-specific, creating significant challenges in multilingual contexts, particularly for less widely-spoken languages (Névéol et al., 2018).

Certain settings such as multi-centric clinical trials, large multinational clinical studies and trials, rare disease characterization, comparative analysis between clinical sites, and international medical collaborations and clinical research require comparable annotation criteria and information extraction systems across languages and clinical content written in different languages. This goal is difficult to achieve due to the scarcity of multilingual clinical corpora annotated under consistent or comparable data labeling criteria and well defined annotation guidelines and criteria.

Recent advances in machine translation, as well as LLMs and generative AI, offer promising opportunities for creating annotated corpora and applying NLP technologies across multiple languages (Gilardi et al., 2023). MT systems or LLMs can translate annotated corpora from one language to another while preserving the integrity and contextual meaning of the original annotations. The use of annotation projection, advanced entity alignment strategies, as well as LLMs for data annotation, can foster the development of comparable multilingual clinical corpora, which in turn can serve as training resources for advanced comparable multilingual clinical entity recognition techniques (Ni et al., 2017; Politov et al., 2025).

In this context, the MultiClinAI (Multilingual Clinical Entity Annotation Projection and Extraction) shared task aims at the automatic generation of comparable multilingual clinical corpora. MultiClinAI focuses on three key clinical entities of high relevance for biomedical data analysis, predictive modeling applications, and the development and evaluation of multilingual clinical NER solutions (diseases, symptoms and procedures) in seven Eu-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 309–331

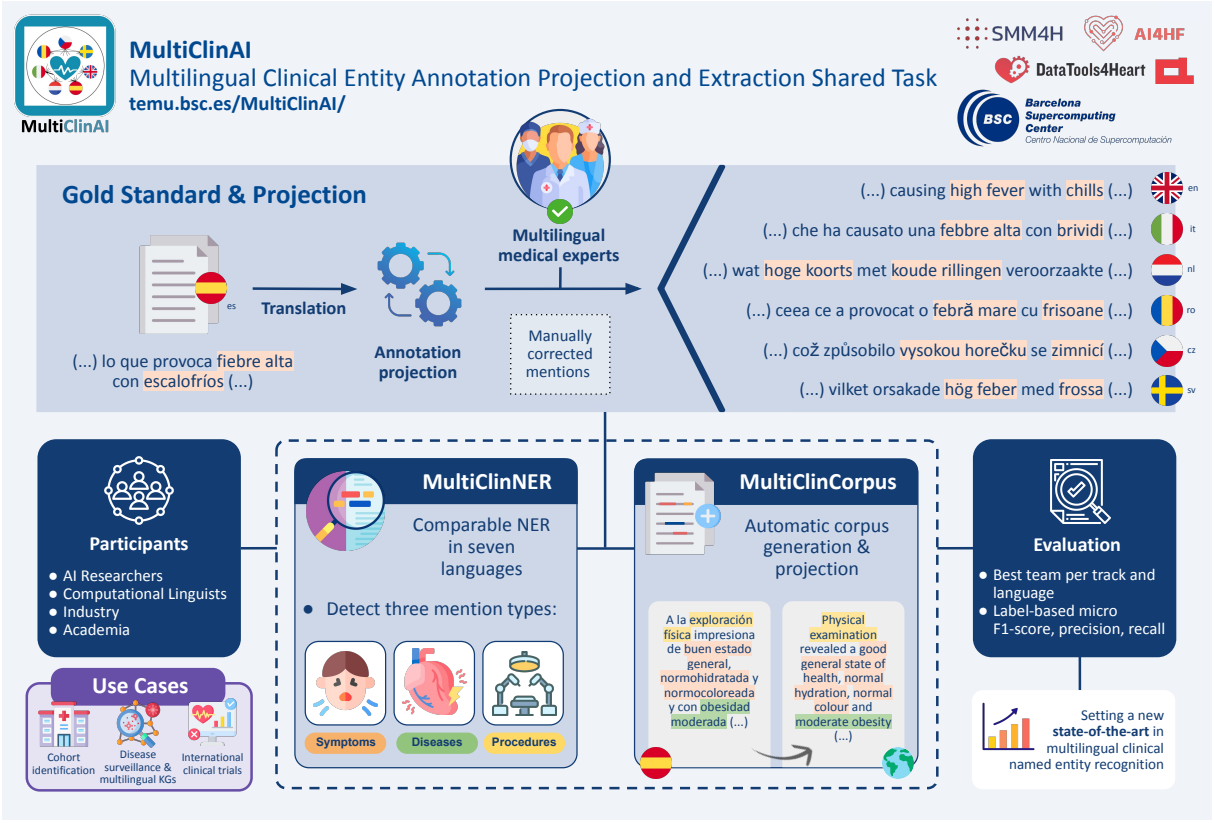

Figure 1: Overview figure for the MultiClinAI shared task, originally used for dissemination purposes.

ropean languages covering different language families: Czech, Dutch, English, Italian, Romanian, Spanish and Swedish.

2 Task Description

2.1 Shared task description

MultiClinAI was held as part of the #SMM4H-HeaRD workshop (Lopez-Garcia et al., 2026) at ACL 2026. It explores automatic annotation projection strategies for multilingual clinical corpus creation, as well as comparable multilingual concept extraction solutions, through two different subtracks:

- **MultiClinNER Subtask:** focuses on the evaluation of named entity recognition systems in seven different languages (Czech, Dutch, English, Italian, Romanian, Spanish and Swedish).

- **MultiClinCorpus Subtask:** covers the automatic generation of comparable multilingual corpora through annotation projection. Given a collection of annotated clinical cases in Spanish, participant teams have to return the equivalent entity mentions in the six target languages.

Both subtasks cover the same three entity types: diseases, symptoms and findings, and procedures. Participants were allowed to submit results for any language and label they wished. Figure 1 shows a visual overview of the task setting.

2.2 Evaluation metrics

For a comprehensive evaluation of the different MultiClinAI submissions, we report both strict and character-based precision, recall, and F-1 score. These two evaluation settings provide complementary perspectives on system performance.

The primary evaluation setting is the *strict* criterion, which is also the most informative, as it measures whether the system reconstructs each entity precisely. Under this criterion, a predicted entity is considered correct only if both its span boundaries and label exactly match those of a gold entity. As a result, strict evaluation is highly demanding: even omitting a single word or failing to include the final character of an entity span is penalized as an error. This makes it especially suitable for assessing a system's ability to produce exact extractions.

Language	SM									FZY								
	Disease			*Procedure*			*Symptom*			*Disease*			*Procedure*			*Symptom*		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Czech	0.256	0.135	0.177	0.221	0.187	0.203	0.273	0.141	0.186	0.178	0.161	0.170	0.168	0.211	0.187	0.175	0.161	0.168
English	0.225	0.164	0.190	0.099	0.190	0.131	0.165	0.127	0.143	0.159	0.185	0.171	0.091	0.211	0.127	0.113	0.151	0.129
Spanish	0.196	0.141	0.164	0.250	0.208	0.227	0.243	0.071	0.110	0.154	0.153	0.153	0.185	0.221	0.202	0.159	0.077	0.104
Italian	0.144	0.133	0.138	0.153	0.143	0.148	0.229	0.145	0.178	0.119	0.144	0.130	0.115	0.155	0.132	0.155	0.155	0.155
Dutch	0.311	0.213	0.253	0.131	0.273	0.177	0.212	0.181	0.195	0.225	0.226	0.226	0.117	0.294	0.168	0.146	0.189	0.165
Romanian	0.126	0.138	0.132	0.229	0.180	0.202	0.201	0.137	0.163	0.116	0.172	0.139	0.175	0.207	0.190	0.133	0.158	0.145
Swedish	0.400	0.210	0.248	0.313	0.260	0.284	0.302	0.168	0.216	0.228	0.210	0.219	0.254	0.289	0.271	0.202	0.185	0.193

Table 1: Baseline results for the MultiClinNER task using string matching (SM) and fuzzy matching (FZY). Results are reported separately for Disease, Procedure, and Symptom entities under strict evaluation.

Strict precision, recall, and F-1 score are defined as follows:

$$P_s = \frac{TP}{TP + FP} \qquad R_s = \frac{TP}{TP + FN}$$

$$F1_s = \begin{cases} \frac{2P_sR_s}{P_s+R_s}, & \text{if } P_s + R_s > 0 \\ 0, & \text{otherwise} \end{cases}$$

Additionally, we report a less strict *character-based* evaluation. This alternative is designed to capture partial matches between predicted and gold spans and is therefore less penalizing when the system identifies most of an entity correctly but makes a minor boundary error. In this way, character-based evaluation complements the strict metric by reflecting whether a prediction is substantially correct, even if not perfectly delimited. To calculate character-based precision, recall and F-1 scores, let G be the set of gold entities and P the set of predicted entities. For a gold span $g \in G$ and a predicted span $p \in P$, we define their character overlap as:

$$O(g, p) = |g \cap p|$$

where $O(g, p)$ is the number of overlapping characters between the gold span g and the predicted span p. Based on this overlap, we define the following similarity score:

$$\phi(g, p) = \frac{2\,O(g, p)}{|g| + |p|}$$

Character-based Precision and Recall are then computed by assigning each predicted span to its best overlapping gold span, and each gold span to its best overlapping predicted span:

$$P_c = \frac{1}{|P|} \sum_{p \in P} \max_{g \in G} \phi(g, p)$$

$$R_c = \frac{1}{|G|} \sum_{g \in G} \max_{p \in P} \phi(g, p)$$

Finally, the character-based F1 score is defined as:

$$F1_c = \begin{cases} \frac{2P_cR_c}{P_c+R_c}, & \text{if } P_c + R_c > 0 \\ 0, & \text{otherwise} \end{cases}$$

An official evaluation library was also made publicly available on GitHub during the competition[1].

2.3 Baseline

To provide a reference point for system comparison, we implemented two simple baseline approaches based on cross-lingual lexical matching. Following a vocabulary transfer strategy, we first constructed gazetteers of annotated entity mentions (Disease, Symptom, and Procedure) from the training data. These vocabularies were then directly applied to the test sets of each target language.

The first baseline (**SM**) consists of exact string matching, where entity mentions from the gazetteer are searched verbatim in the test documents. The second baseline (**FZY**) use fuzzy matching, allowing approximate string matches to account for minor lexical variations.

The results obtained with these baseline systems are reported in Table 1. This setup provides a lightweight and language-agnostic baseline, enabling the assessment of how far more advanced systems improve over simple lexical transfer methods.

3 Corpus and Resources

The MultiClinNER and MultiClinCorpus datasets are a collection of clinical case reports from multiple clinical specialties annotated with diseases, symptoms and procedures in seven different languages (Czech, English, Dutch, Italian, Romanian, Spanish and Swedish). The datasets integrate previously released data in Spanish, namely DisTEMIST

[1] https://github.com/nlp4bia-bsc/MultiClinAIEval

(Miranda-Escalada et al., 2022), SympTEMIST (Lima-López et al., 2023b), MedProcNER (Lima-López et al., 2023a) and CardioCCC (Lima-López et al., 2024), expanding them with new clinical cases, annotations and languages. Table 2 shows a general overview of the two corpora in all 7 languages of the task, while Tables 6 and 7 in Appendix A show a more detailed breakdown of the number of annotations in each dataset. Notably, the content for both datasets is the mostly the same, but we introduce them as two separate entities since the distribution of the documents is not exactly the same for both, as well as to highlight their distinct purposes within the task (named entity recognition and annotation projection). A major difference between both datasets is their test set. While the test set for the MultiClinCorpus is the same across all languages, the MultiClinNER test set includes additional clinical cases, making its size different for each language.

The clinical cases in the corpus were annotated originally in Spanish by experienced clinical domain experts using the guidelines specially developed for the task[2]. The multilingual versions were then developed using annotation projection and human validation using the Spanish annotations as a seed version. First, both the texts and annotations were automatically translated into each of the target languages (Czech, English, Dutch, Italian, Romanian and Swedish). The translated annotations were then mapped onto the translated texts using a lexical lookup. The resulting dataset was subsequently manually revised by bilingual clinical experts to make sure that each target version resembled as closely as possible its Spanish counterpart. This was done using the annotation tool brat (Stenetorp et al., 2012)'s side-by-side view, which allowed annotators to see both versions of the document at the same time. To make up for possible mistranslations by the machine translation system, annotators were instructed to provide alternative translations for any mistranslated clinical entities. These were then introduced in the text in a post-processing step. For this reason, the text files for each entity type (disease, symptom, procedure) are different.

To compensate for the use of machine-translated clinical cases, a subset of native clinical case reports for each language was also included. For these corpora, the native clinical cases in each language were translated into Spanish. Manual annotation was then performed on the Spanish versions of the texts by the same annotators who worked on the SpaCCC and CardioCCC corpora. Once this step was completed, the annotations were translated and mapped back into the original language and subsequently revised by bilingual experts. These additional clinical cases are part of the MultiClinNER test set, which explains why the number of documents differs across languages and is considerably higher in Spanish.

Additionally, a large collection of clinical case reports was released together with the MultiClinAI corpora to serve as a background set, for which participants were asked to generate predictions alongside the test set. This background set was intended to examine the scalability of participating systems and discourage manual correction or revision of submissions. The resulting set of participant predictions for these files will be harmonized and released as a Silver Standard as an additional resource for the task.

All MultiClinAI data and resources are publicly available on Zenodo[3].

4 Results

4.1 Participation Overview

A total of 56 teams registered for the MultiClinAI shared task, of which 21 submitted at least one run, representing 13 different countries. Table 3 provides an overview of the participating teams. In total, the task received 531 final system submissions, indicating a substantial level of participation. A large proportion of these submissions (471) were associated with the MultiClinNER subtask, whereas 60 submissions were made to the MultiClinCorpus track. Only limited overlap between subtasks was observed, with two teams, Parallia and blue, contributing systems to both.

4.2 Submissions Results

Given the large number of submissions received in the MultiClinAI shared task, we report only the best-performing result for each entity type and language combination in the main results table. This allows for a clearer and more concise comparison of system performance across the different evaluation settings. The complete set for subtasks Multi-

[2]Available on Zenodo in Spanish (https://doi.org/10.5281/zenodo.10171539) and English (https://doi.org/10.5281/zenodo.10171646).

[3]https://zenodo.org/records/19334278

MultiClinNER					MultiClinCorpus				
Lang	**Split**	**Docs**	**Tokens**	**Anns**	**Lang**	**Split**	**Docs**	**Tokens**	**Anns**
cz	Train	1,258	784,752	81,100	cz	Train	1,258	784,752	81,100
en	Train	1,258	880,670	79,316	en	Train	1,258	880,670	79,316
es	Train	1,258	873,603	83,507	es	Train	1,258	873,603	83,507
it	Train	1,258	912,174	81,482	it	Train	1,258	912,174	81,482
nl	Train	1,258	812,110	80,853	nl	Train	1,258	812,110	80,853
ro	Train	1,258	871,469	79,889	ro	Train	1,258	871,469	79,889
sv	Train	1,258	776,893	80,190	sv	Train	1,258	776,893	80,190
cz	Test	337	123,923	13,121	cz	Test	250	86,154	9,257
en	Test	600	327,829	28,905	en	Test	250	100,827	9,215
es	Test	1,954	843,103	70,771	es	Test	250	101,152	9,322
it	Test	567	224,369	18,259	it	Test	250	104,811	9,209
nl	Test	600	163,748	15,183	nl	Test	250	94,141	9,334
ro	Test	363	131,312	12,383	ro	Test	250	100,418	9,228
sv	Test	350	128,179	13,242	sv	Test	250	87,707	9,237
Total		13,577	7,854,134	738,201	**Total**		10,556	6,586,881	631,139

Table 2: General statistics for the MultiClinAI datasets. 'cz' stands for Czech, 'en' for English, 'es' for Spanish, 'it' for Italian, 'nl' for Dutch, 'ro' for Romanian and 'sv' for Swedish.

ClinNER and MultiClinCorpus are shown, respectively, on Tables 8 and 9 as part of Appendix B.

The results obtained for the MultiClinNER subtask are summarized in Table 4, which reports the best-performing systems for each entity type and language under strict evaluation. Overall, the results show a consistent performance across languages, with F1-scores generally ranging between 0.64 and 0.82 depending on the entity type and target language. It is worth highlighting that lower performance values are mainly associated with the symptom entity type, which is typically more challenging due to its higher variability and descriptive nature, as well as with languages that are comparatively less resourced and documented.

Among the evaluated entities, disease mentions tend to achieve the highest performance across most languages, with peak results observed in Spanish (0.824) and English (0.805). Procedure extraction also shows competitive performance, particularly in Spanish (0.813), while symptom recognition remains the most challenging entity type, yielding comparatively lower scores across all languages.

From a multilingual perspective, results indicate that systems generalize reasonably well across languages, although a slight degradation in performance can be observed for certain languages such as Dutch and Czech. Notably, the best-performing system across all evaluated settings was consistently submitted by the `BIT.UA` team, which combined multilingual transformer-based models (e.g., XLM-RoBERTa) with ensemble strategies across multiple runs. This combination of cross-lingual modeling and ensembling techniques appears to be particularly effective, demonstrating strong robustness and adaptability in multilingual clinical entity recognition scenarios.

In addition to the MultiClinNER subtask, MultiClinAI also introduced the MultiClinCorpus track, which focuses on annotation projection. This subtask represents a novel evaluation setting aimed at exploring automatic strategies to extend high-quality clinical annotations from well-resourced languages to less-resourced ones. By leveraging cross-lingual transfer techniques, MultiClinCorpus seeks to facilitate the creation of comparable multilingual clinical corpora with reduced manual annotation effort.

The results obtained for the MultiClinCorpus subtask are presented in Table 5. Performance levels are notably higher than in the MultiClinNER task, with F1-scores consistently above 0.77 and reaching up to 0.90 in several language-entity combinations. This reflects the comparatively more constrained nature of the annotation projection task, where systems benefit from aligned source annotations.

Across all evaluated settings, the `Parallia` team achieved the best results. Their approach, based

Team Name	Affiliation	Country	Subtasks	Citation
blue	Thapar Institute of Engineering and Technology	India	NER & Corpus	(Sharma et al., 2026)
BIT.UA	Universidade de Aveiro	Portugal	NER	(Jonker and Matos, 2026)
Discovery@FI	Masaryk University	Czechia	NER	(Zelina and Novacek, 2026)
Dr-BERT-NL	University Medical Centre Groningen	Netherlands	NER	(Danoe et al., 2026)
DT4H_NL	University Medical Center Utrecht	Netherlands	NER	(van Es, 2026)
Enigma	Sofia University St. Kliment Ohridski	Bulgaria	NER	(Vassileva et al., 2026)
HSE NLP	Higher School of Economics	Germany	NER	-
ICB-UMA	University of Malaga	Spain	Corpus	(Rey-Blanes et al., 2026)
IIC-UC3M	Universidad Carlos III de Madrid	Spain	NER	-
LemonHeard	Sapienza University of Rome	Italy	NER	-
Livia Clarete	The City University of New York	United States	NER	-
Lotus Orchid	Vrije Universiteit Amsterdam	Netherlands	NER	(Arnoult et al., 2026)
LSI_UNED	Universidad Nacional de Educación a Distancia	Spain	NER	(Ramirez-Arrabe et al., 2026)
NLP-FBK	Fondazione Bruno Kessler	Italy	NER	-
Parallia	Parallia AI4Health	Belgium	NER & Corpus	(Remy, 2026)
SIEMENS	Siemens S.R.L. - Transilvania University of Brasov	Romania	NER	(Danu, 2026)
SINAI	University of Jaén	Spain	NER	(Molino Piñar et al., 2026)
SuSh	Utica University	United States	NER	-
Sycamore	University College London	United Kingdom	NER	-
Unibo-NLP	University of Bologna	Italy	NER	-
Vinland Vector	Chittagong University of Engineering & Technology	Bangladesh	NER	(Das et al., 2026)

Table 3: Overview of participant teams in the MultiClinAI shared task.

on the *ClinicalAligner* system, demonstrates strong effectiveness in transferring annotations across languages, achieving particularly high scores for disease entities, with peak performance observed in Romanian (0.898) and English (0.896). Notably, the system also maintains competitive performance in less-resourced languages such as Swedish (0.828 for diseases) and Czech (0.851), highlighting the robustness of annotation projection techniques even in settings with more limited linguistic resources. Beyond F1-score, the system also exhibits consistently high precision and recall values across all settings, generally remaining above 0.80, which reflects both accurate and well-balanced predictions.

In summary, the results obtained across the MultiClinNER and MultiClinCorpus subtasks provide complementary insights into multilingual clinical information extraction. The MultiClinNER results highlight the challenges associated with direct entity recognition in a multilingual setting, particularly for more complex and variable entity types such as symptoms. Nevertheless, the use of multilingual transformer-based architectures, especially when combined with ensemble strategies, enables strong and consistent performance across languages. In contrast, the MultiClinCorpus subtask demonstrates that annotation projection constitutes a highly effective approach for extending annotated clinical resources to multiple languages, achieving high precision and recall even in less-resourced scenarios. These findings indicate that the integration of direct multilingual modeling and annotation projection strategies represents a robust and scalable framework for the development of clinical NLP systems across languages.

4.3 Methodologies

Participants in the MultiClinNER subtask explored a wide range of modeling strategies, with the majority of systems relying on transformer-based architectures. A common baseline approach consisted of fine-tuning multilingual models such as XLM-RoBERTa for token classification, as adopted by several teams. In contrast, other participants focused on domain-adapted and language-specific models, leveraging pretrained representations tailored to biomedical or clinical text, including models such as RigoBERTa, MedRoBERTa.nl, and MrBERT-biomed. Some teams further specialized their systems by training separate models per language or entity type, highlighting the importance of domain and linguistic adaptation in multilingual clinical NER.

Beyond standard fine-tuning, a number of teams

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
	Czech					**English**				
Disease	BIT.UA	ensemble-xlm-roberta	0.740	0.710	0.725	BIT.UA	ensemble-all-runs	0.816	0.795	0.805
Procedure	BIT.UA	ensemble-xlm-roberta	0.748	0.708	0.727	BIT.UA	ensemble-all-runs	0.792	0.718	0.753
Symptom	BIT.UA	ensemble-xlm-roberta	0.692	0.651	0.671	BIT.UA	ensemble-all-runs	0.774	0.711	0.741
	Spanish					**Italian**				
Disease	BIT.UA	ensemble-all-runs	0.810	0.839	0.824	BIT.UA	ensemble-xlm-roberta	0.803	0.702	0.749
Procedure	BIT.UA	ensemble-all-runs	0.814	0.813	0.813	BIT.UA	ensemble-xlm-roberta	0.749	0.728	0.739
Symptom	BIT.UA	ensemble-all-runs	0.784	0.774	0.779	BIT.UA	ensemble-xlm-roberta	0.734	0.648	0.689
	Dutch					**Romanian**				
Disease	BIT.UA	ensemble-xlm-roberta	0.736	0.739	0.737	BIT.UA	ensemble-xlm-roberta	0.779	0.762	0.770
Procedure	BIT.UA	ensemble-xlm-roberta	0.723	0.714	0.718	BIT.UA	ensemble-xlm-roberta	0.765	0.740	0.752
Symptom	BIT.UA	ensemble-xlm-roberta	0.663	0.633	0.647	BIT.UA	ensemble-xlm-roberta	0.768	0.650	0.704
	Swedish									
Disease	BIT.UA	ensemble-xlm-roberta	0.759	0.717	0.738	-	-	-	-	-
Procedure	BIT.UA	ensemble-xlm-roberta	0.744	0.732	0.738	-	-	-	-	-
Symptom	BIT.UA	ensemble-xlm-roberta	0.717	0.669	0.692	-	-	-	-	-

Table 4: Best-performing systems for the MultiClinNER subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R) and F1-score (F1) are computed using exact span and label matching. *Note: For each entity-language pair, only the top-scoring run is shown.*

explored more advanced strategies to improve performance. Ensemble-based approaches were widely used, ranging from homogeneous ensembles of multilingual models to more complex multi-level and heterogeneous combinations integrating multilingual, domain-adapted, and language-specific models. In parallel, several participants investigated alternative formulations of the task, particularly multilabel approaches, as opposed to the traditional multiclass setting, allowing greater flexibility in modeling entity assignments. Additional variations included the use of cross-validation, data augmentation, and post-processing techniques to refine entity boundaries and improve robustness.

Finally, a subset of systems departed from the conventional token classification paradigm. These include zero-shot and generalized NER approaches based on models such as GLiNER, which enable label-driven entity extraction without task-specific fine-tuning, as well as large language model (LLM)-based methods leveraging generative architectures or parameter-efficient fine-tuning (e.g., LoRA). Alternative architectural designs were also explored, such as relation-based models (W2NER) and hybrid Transformer-CNN architectures, demonstrating the diversity of methodological directions considered in this shared task. Overall, the submitted systems reflect a rich landscape of approaches, combining advances in multilingual modeling, domain adaptation, and emerging paradigms in clinical NLP.

For MultiClinCorpus, ICB-UMA addressed Spanish-to-English annotation projection through a candidate-ranking pipeline. Their runs evolved from a heuristic method based on token-length constraints, character-offset proximity, and surface similarity metrics, to a supervised XGBoost-based ranker using surface, positional, and semantic features. In the later submissions, this pipeline was further complemented with an LLM-as-judge correction step to revise uncertain projections. In contrast, Blue adopted a lightweight Spanish-to-target projection strategy based on cognate detection and fuzzy string matching over the Spanish annotations, leveraging surface similarity between related terms across languages.

Parallia, which achieved the best performance across all settings, adopted a cross-lingual token alignment approach for Spanish-to-target annotation projection. Their method relies on a pretrained ClinicalAligner model to map token-level embeddings between languages using optimal transport, enabling direct span projection. This approach was further enhanced by incorporating predictions from their MultiClinNER system and, in some runs, minimal fine-tuning on the task data, resulting in a ro-

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
	Czech					**English**				
Disease	Parallia	ClinicalAligner26AM-S3	0.845	0.856	0.851	Parallia	ClinicalAligner26AM-S3	0.891	0.901	0.896
Procedure	Parallia	ClinicalAligner26AM-S3	0.817	0.820	0.819	Parallia	ClinicalAligner26AM-S3	0.835	0.847	0.841
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.876	0.882	0.879
	Swedish					**Italian**				
Disease	Parallia	ClinicalAligner26AM-S2	0.822	0.834	0.828	Parallia	ClinicalAligner26AM-S3	0.876	0.887	0.882
Procedure	Parallia	ClinicalAligner26AM-S3	0.803	0.809	0.806	Parallia	ClinicalAligner26AM-S3	0.792	0.806	0.799
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.829	0.833	0.831
	Dutch					**Romanian**				
Disease	Parallia	ClinicalAligner26AM-S3	0.818	0.822	0.820	Parallia	ClinicalAligner26AM-S3	0.894	0.902	0.898
Procedure	Parallia	ClinicalAligner26AM-S3	0.774	0.766	0.770	Parallia	ClinicalAligner26AM-S3	0.850	0.861	0.856
Symptom	Parallia	ClinicalAligner26AM-S3	0.767	0.769	0.768	Parallia	ClinicalAligner26AM-S3	0.858	0.864	0.861

Table 5: Best-performing systems for the MultiClinCorpus subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R), and F1-score (F1) are computed using exact span and label matching after annotation projection. *Note: For each entity-language pair, only the top-scoring run is shown.*

bust and highly effective projection pipeline across languages.

5 Discussion

The MultiClinAI task can be considered an effort to foster more collaborative scenarios for generating clinical corpora and NLP resources beyond language-specific silos. It is critical to more efficiently exploit the still limited number of high-quality, manually annotated clinical corpora, not only for a single language but also for potential extension under cross-lingual adaptation scenarios. This is particularly relevant in clinical and public health contexts where information needs to be extracted consistently across languages, for example to support infectious disease surveillance and the characterization of diseases, symptoms, procedures, and comorbidities from multilingual clinical text. Large European projects such as the cardiovascular AI consortia DataTools4Heart and AI4HF require access to comparable multilingual NLP systems as part of advanced predictive disease modelling. MultiClinAI and the systems implemented by participants can potentially be regarded as contributions in this direction. However, extending coverage to additional languages such as German, French, and Chinese, and in particular to under-resourced languages, remains a pressing need for enabling globally inclusive clinical language technology infrastructures.

Acknowledgments

MultiClinAI was funded by the European projects DataTools4Heart (Grant Agreement No. 101057849) and AI4HF (Grant Agreement No. 101080430). Fernando Gallego Donoso and Salvador Lima-López fellowship within the "Generación D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1). Funded by the European Union NextGenerationEU funds, through PRTR.

References

Sophie Arnoult, Shutao Chen, and Piek Vossen. 2026. LotusOrchid at SMM4H–HeaRD 2026: Fitting pretrained encoders for Dutch medical data. 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.

Gijs Danoe, Matthijs S. Berends, Andreas Voss, and Axel Hamprecht. 2026. Dr-BERT-NL at SMM4H–HeaRD 2026: DOKTERBERT – Ontology-Grounded Contextual Representations for Dutch Clinical NLP. 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.

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

Nirjhar Das, Rathijit Aich, and Mahfuzulhoq Chowdhury. 2026. Vinland_Vector at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Query-Augmented Clinical NER for English. 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.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Richard Adolph Aires Jonker and Sérgio Matos. 2026. BIT.UA at SMM4H–HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles. 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.

S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, and M. Krallinger. 2023a. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In *Working Notes of CLEF*.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023b. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Anoop Shah, Anastasios Nentidis, and 1 others. 2024. Overview of MultiCardioNER task at BioASQ 2024 on Medical Specialty and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In *Working Notes of CLEF*.

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.

A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, and M. Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *CLEF (Working Notes)*, pages 179–203.

Lucas Molino Piñar, Manuel Carlos Diaz-Galiano, and María-Teresa Martín-Valdivia. 2026. SINAI at MultiClinAI 2026: Multilingual Clinical NER with MrBERT-biomed and Optuna Hyperparameter Optimization. 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.

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.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Andrei Politov, Oleh Shkalikov, Rene Jäkel, and Michael Färber. 2025. Revisiting projection-based data transfer for cross-lingual named entity recognition in low-resource languages. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 499–507, Tallinn, Estonia. University of Tartu Library.

Alicia Ramirez-Arrabe, Juan Martinez-Romo, and Andrés Duque. 2026. LSI_UNED at MultiClinAI 2026: Grid-Based Biomedical Named Entity Recognition Across Languages and Entity Types. 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.

François Remy. 2026. Parallia at SMM4H-HeaRD 2026: ClinicalAligner26AM: A Cross-Lingual Aligner for Dataset Translation; Evidences from the MultiClinCorpus Shared Task. 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.

Alvaro Rey-Blanes, Sara Giménez-Gómez, Francisco J. Veredas, and Francisco J. Moreno-Barea. 2026. ICB-UMA at SMM4H–HeaRD 2026: Hybrid Clinical Entity Projection for MultiClinAI: Adaptive Candidate Windows, XGBoost, and LLM Refinement. 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.

Krish Sharma, Rhea Singhal, and Jatin Bedi. 2026. blue at SMM4H-HeaRD 2026: Class-Weighted Transformer Ensembles with Structured Decoding and Chain-of-Thought Blending across Six Health NLP Shared Tasks. 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

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

Sylvia Vassileva, Plamena Ilieva, Teodor Kostadinov, Monika Peteva Petkova, Daniel Manchevski, Vitosh Doynov, Ivan Koychev, and Svetla Boytcheva. 2026. Enigma at SMM4H–HeaRD 2026: Leveraging Multilingual Pre-trained Models for Clinical Named Entity Recognition. 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.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of biomedical informatics*, 77:34–49.

Petr Zelina and Vit Novacek. 2026. Discovery@FI at SMM4H–HeaRD 2026: Ensemble Character Classifier for Multilingual Clinical NER. 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.

A Detailed corpus statistics

Lang	Split	Docs	Tokens	SYMPTOM	PROCEDURE	DISEASE	Total
cz	Train	1,258	784,752	27,806	27,501	25,793	81,100
en	Train	1,258	880,670	27,465	26,733	25,118	79,316
es	Train	1,258	873,603	29,074	28,137	26,296	83,507
it	Train	1,258	912,174	27,928	27,394	26,159	81,482
nl	Train	1,258	812,110	27,675	27,445	25,733	80,853
ro	Train	1,258	871,469	27,015	27,313	25,561	79,889
sv	Train	1,258	776,893	27,531	27,079	25,580	80,190
cz	Test	337	123,923	3,998	5,368	3,755	13,121
en	Test	600	327,829	8,873	10,978	9,054	28,905
es	Test	1,954	843,103	24,449	24,983	21,339	70,771
it	Test	567	224,369	5,860	7,144	5,255	18,259
nl	Test	600	163,748	5,466	5,138	4,579	15,183
ro	Test	363	131,312	3,971	4,574	3,838	12,383
sv	Test	350	128,179	4,683	4,933	3,626	13,242
Total		13,577	7,854,134	251,794	254,720	231,686	738,201

Table 6: Detailed statistics for the MultiClinNER dataset including entity type distribution

Lang	Split	Docs	Tokens	SYMPTOM	PROCEDURE	DISEASE	Total
cz	Train	1,258	784,752	27,806	27,501	25,793	81,100
en	Train	1,258	880,670	27,465	26,733	25,118	79,316
es	Train	1,258	873,603	29,074	28,137	26,296	83,507
it	Train	1,258	912,174	27,928	27,394	26,159	81,482
nl	Train	1,258	812,110	27,675	27,445	25,733	80,853
ro	Train	1,258	871,469	27,015	27,313	25,561	79,889
sv	Train	1,258	776,893	27,531	27,079	25,580	80,190
cz	Test	250	86,154	3,094	3,603	2,560	9,257
en	Test	250	100,827	3,080	3,568	2,567	9,215
es	Test	250	101,152	3,104	3,619	2,599	9,322
it	Test	250	104,811	3,087	3,553	2,569	9,209
nl	Test	250	94,141	3,096	3,654	2,584	9,334
ro	Test	250	100,418	3,080	3,572	2,576	9,228
sv	Test	250	87,707	3,095	3,587	2,555	9,237
Total		10,556	6,586,881	210,130	214,758	206,251	631,139

Table 7: Detailed statistics for the MultiClinCorpus dataset including entity type distribution

B Complete task results

B.1 MultiClinNER

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Language: Czech (cz)							
Entity: Disease							
BIT.UA	ensemble-x...	0.740	0.710	0.725	0.839	0.810	0.825
BIT.UA	xlm-robert...	0.722	0.705	0.713	0.827	0.812	0.820
BIT.UA	ensemble-a...	0.747	0.664	0.703	0.846	0.756	0.799
SIEMENS	cz_disease...	0.713	0.673	0.693	0.815	0.772	0.793
SIEMENS	cz_disease...	0.695	0.677	0.686	0.804	0.786	0.795
BIT.UA	ensemble-n...	0.714	0.656	0.684	0.823	0.755	0.788
Parallia	run	0.656	0.692	0.674	0.778	0.820	0.798
SIEMENS	cz_disease...	0.688	0.642	0.664	0.810	0.756	0.782
SIEMENS	cz_disease...	0.677	0.649	0.663	0.788	0.757	0.772
SIEMENS	cz_disease...	0.728	0.606	0.661	0.824	0.688	0.750
BIT.UA	robeczech-...	0.675	0.647	0.661	0.792	0.762	0.777
Enigma	robeczech-...	0.674	0.637	0.655	0.797	0.752	0.774
Enigma	robeczech-...	0.667	0.632	0.649	0.793	0.750	0.771
Discovery@FI	run	0.610	0.670	0.639	0.741	0.797	0.768
Enigma	xlmr-crf-c...	0.650	0.626	0.638	0.775	0.745	0.759
Enigma	xlmr-para-...	0.642	0.615	0.628	0.766	0.732	0.748
SIEMENS	cz_disease...	0.630	0.626	0.628	0.759	0.756	0.757
Enigma	ensemble_c...	0.565	0.674	0.615	0.692	0.814	0.748
DT4H_NL	EuroBERT61...	0.649	0.520	0.577	0.790	0.635	0.704
DT4H_NL	EuroBERT61...	0.644	0.507	0.567	0.793	0.628	0.701
DT4H_NL	DeBERTa_mu...	0.582	0.514	0.546	0.726	0.646	0.683
blue	run	0.247	0.230	0.238	0.695	0.643	0.668
nlp-fbk	run	0.001	0.001	0.001	0.665	0.559	0.608
Livia_Clarete	*run1*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Livia_Clarete	*run2*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Entity: Procedure							
BIT.UA	ensemble-x...	0.748	0.708	0.727	0.855	0.813	0.834
BIT.UA	ensemble-a...	0.753	0.684	0.717	0.859	0.784	0.820
BIT.UA	xlm-robert...	0.713	0.707	0.710	0.835	0.834	0.835
BIT.UA	ensemble-n...	0.725	0.686	0.705	0.840	0.797	0.818
SIEMENS	cz_procedu...	0.719	0.683	0.701	0.834	0.791	0.812
Parallia	run	0.674	0.703	0.688	0.809	0.847	0.828
SIEMENS	cz_procedu...	0.694	0.680	0.687	0.824	0.805	0.814
BIT.UA	robeczech-...	0.679	0.670	0.675	0.812	0.806	0.809
SIEMENS	cz_procedu...	0.727	0.621	0.669	0.844	0.722	0.778
SIEMENS	cz_procedu...	0.681	0.656	0.668	0.821	0.795	0.808
SIEMENS	cz_procedu...	0.696	0.642	0.668	0.827	0.766	0.796
Enigma	robeczech-...	0.664	0.660	0.662	0.802	0.792	0.797
Discovery@FI	run	0.629	0.690	0.658	0.771	0.841	0.805
Enigma	xlmr-crf-c...	0.657	0.654	0.655	0.795	0.791	0.793

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Enigma	ensemble_c...	0.617	0.685	0.649	0.758	0.832	0.793
SIEMENS	cz_procedu...	0.659	0.639	0.649	0.808	0.787	0.797
DT4H_NL	EuroBERT61...	0.672	0.550	0.605	0.816	0.675	0.739
DT4H_NL	EuroBERT61...	0.687	0.532	0.600	0.833	0.651	0.731
blue	run	0.347	0.333	0.340	0.747	0.717	0.732
nlp-fbk	run	0.000	0.000	0.000	0.650	0.499	0.565
Entity: Symptom							
BIT.UA	ensemble-8...	0.692	0.651	0.671	0.805	0.757	0.780
BIT.UA	xlm-robert...	0.657	0.661	0.659	0.774	0.781	0.777
BIT.UA	ensemble-2...	0.695	0.614	0.652	0.803	0.710	0.754
BIT.UA	ensemble-2...	0.658	0.616	0.636	0.775	0.723	0.748
SIEMENS	cz_symptom...	0.649	0.613	0.631	0.771	0.726	0.748
SIEMENS	cz_symptom...	0.636	0.603	0.619	0.770	0.728	0.748
SIEMENS	cz_symptom...	0.623	0.613	0.618	0.747	0.735	0.741
Enigma	robeczech-...	0.624	0.591	0.607	0.748	0.708	0.728
BIT.UA	robeczech-...	0.609	0.604	0.606	0.741	0.736	0.738
SIEMENS	cz_symptom...	0.609	0.586	0.597	0.735	0.707	0.721
Enigma	xlmr-crf-c...	0.610	0.583	0.596	0.738	0.706	0.721
Enigma	ensemble_c...	0.564	0.629	0.595	0.689	0.765	0.725
Parallia	run	0.550	0.625	0.585	0.694	0.783	0.736
Enigma	xlmr-os1-c...	0.594	0.570	0.582	0.725	0.694	0.709
SIEMENS	cz_symptom...	0.697	0.477	0.567	0.808	0.554	0.657
SIEMENS	cz_symptom...	0.543	0.588	0.564	0.689	0.741	0.714
DT4H_NL	EuroBERT61...	0.625	0.493	0.551	0.771	0.609	0.681
DT4H_NL	EuroBERT61...	0.635	0.479	0.546	0.773	0.586	0.666
Discovery@FI	run	0.494	0.600	0.542	0.655	0.767	0.707
blue	run	0.272	0.228	0.248	0.709	0.595	0.647
nlp-fbk	run	0.000	0.000	0.000	0.485	0.278	0.353
Language: English (en)							
Entity: Disease							
BIT.UA	ensemble-a...	0.816	0.795	0.805	0.884	0.865	0.874
BIT.UA	ensemble-x...	0.805	0.794	0.799	0.877	0.868	0.872
BIT.UA	ensemble-n...	0.798	0.791	0.794	0.874	0.866	0.870
BIT.UA	xlm-robert...	0.797	0.787	0.792	0.873	0.867	0.870
SIEMENS	en_disease...	0.811	0.754	0.781	0.879	0.817	0.847
LSI_UNED	conf1_biob...	0.763	0.791	0.777	0.842	0.852	0.847
BIT.UA	microsoft-...	0.778	0.776	0.777	0.861	0.863	0.862
SIEMENS	en_disease...	0.792	0.754	0.773	0.868	0.830	0.849
SIEMENS	en_disease...	0.786	0.737	0.761	0.863	0.813	0.838
SIEMENS	en_disease...	0.771	0.750	0.760	0.852	0.832	0.842
Vinland Vector	glimer_lar...	0.746	0.773	0.759	0.821	0.845	0.833
LSI_UNED	conf1_clin...	0.757	0.757	0.757	0.840	0.823	0.831
Parallia	run	0.731	0.774	0.752	0.828	0.875	0.851
Enigma	en_disease	0.742	0.749	0.745	0.838	0.844	0.841
Vinland Vector	gliner_dis...	0.750	0.738	0.744	0.825	0.810	0.818
Enigma	ensemble-e...	0.726	0.749	0.738	0.820	0.844	0.832

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Discovery@FI	run	0.719	0.755	0.737	0.820	0.855	0.837
SIEMENS	en_disease...	0.701	0.761	0.730	0.794	0.864	0.828
Sycamore	run	0.688	0.720	0.703	0.797	0.826	0.811
DT4H_NL	DeBERTa_mu...	0.722	0.681	0.701	0.817	0.774	0.795
DT4H_NL	EuroBERT61...	0.728	0.631	0.676	0.832	0.724	0.774
DT4H_NL	EuroBERT61...	0.745	0.617	0.675	0.848	0.707	0.771
Unibo-NLP	medgemma-e...	0.719	0.586	0.646	0.835	0.660	0.737
Vinland Vector	pubmedbert...	0.773	0.471	0.585	0.873	0.526	0.656
SuSh	run	0.558	0.538	0.548	0.664	0.631	0.647
blue	run	0.369	0.355	0.362	0.812	0.778	0.795
Enigma	gliner-en-...	0.699	0.097	0.170	0.779	0.107	0.189
Unibo-NLP	medgemma-e...	0.087	0.082	0.084	0.230	0.109	0.147
nlp-fbk	run	0.000	0.000	0.000	0.695	0.635	0.664
HSE NLP	run	0.000	0.000	0.000	0.104	0.000	0.001
Livia_Clarete	*run1*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Livia_Clarete	*run2*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Livia_Clarete	*run3*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Livia_Clarete	*run4*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Entity: Procedure							
BIT.UA	ensemble-a...	0.792	0.718	0.753	0.885	0.804	0.843
BIT.UA	ensemble-x...	0.767	0.734	0.750	0.870	0.835	0.853
LSI_UNED	conf1_biob...	0.758	0.729	0.744	0.860	0.807	0.833
BIT.UA	ensemble-n...	0.772	0.713	0.741	0.875	0.808	0.840
BIT.UA	xlm-robert...	0.750	0.728	0.739	0.861	0.840	0.850
SIEMENS	en_procedu...	0.758	0.702	0.729	0.858	0.796	0.826
SIEMENS	en_procedu...	0.743	0.713	0.728	0.849	0.813	0.831
Vinland Vector	run	0.769	0.688	0.726	0.851	0.756	0.801
SIEMENS	en_procedu...	0.760	0.694	0.725	0.859	0.782	0.819
LSI_UNED	conf1_clin...	0.753	0.698	0.724	0.852	0.774	0.811
SIEMENS	en_procedu...	0.734	0.712	0.723	0.846	0.819	0.832
BIT.UA	microsoft-...	0.749	0.693	0.720	0.863	0.801	0.831
Parallia	run	0.698	0.725	0.711	0.820	0.850	0.835
Enigma	run	0.701	0.703	0.702	0.830	0.830	0.830
SIEMENS	en_procedu...	0.694	0.700	0.697	0.815	0.822	0.819
Discovery@FI	run	0.665	0.718	0.690	0.797	0.848	0.822
DT4H_NL	EuroBERT61...	0.720	0.586	0.646	0.846	0.692	0.761
DT4H_NL	EuroBERT61...	0.735	0.576	0.645	0.862	0.678	0.759
Sycamore	run	0.632	0.629	0.631	0.781	0.770	0.775
SuSh	run	0.519	0.566	0.542	0.634	0.678	0.655
blue	run	0.419	0.406	0.412	0.780	0.755	0.767
nlp-fbk	run	0.001	0.001	0.001	0.688	0.626	0.656
HSE NLP	run	0.005	0.000	0.000	0.197	0.004	0.008
Entity: Symptom							
BIT.UA	ensemble-a...	0.774	0.711	0.741	0.848	0.780	0.812
BIT.UA	ensemble-x...	0.762	0.712	0.736	0.842	0.786	0.813
BIT.UA	ensemble-n...	0.757	0.713	0.734	0.837	0.789	0.812
BIT.UA	xlm-robert...	0.749	0.703	0.725	0.836	0.785	0.809

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
BIT.UA	microsoft-...	0.738	0.691	0.714	0.827	0.776	0.801
LSI_UNED	conf1_biob...	0.743	0.680	0.710	0.831	0.741	0.783
SIEMENS	en_symptom...	0.742	0.667	0.703	0.824	0.741	0.780
SIEMENS	en_symptom...	0.721	0.679	0.699	0.811	0.764	0.787
LSI_UNED	conf1_clin...	0.740	0.661	0.698	0.827	0.722	0.771
SIEMENS	en_symptom...	0.706	0.668	0.686	0.798	0.754	0.775
SIEMENS	en_symptom...	0.706	0.666	0.685	0.802	0.755	0.778
Vinland Vector	gliner_0.5...	0.646	0.704	0.673	0.729	0.787	0.757
Vinland Vector	gliner_0.6...	0.773	0.586	0.667	0.836	0.631	0.719
Enigma	run	0.666	0.662	0.664	0.786	0.776	0.781
SIEMENS	en_symptom...	0.655	0.669	0.662	0.762	0.775	0.768
Parallia	run	0.636	0.674	0.654	0.760	0.799	0.779
Sycamore	run	0.653	0.627	0.640	0.775	0.736	0.755
Discovery@FI	run	0.604	0.652	0.627	0.739	0.781	0.759
DT4H_NL	EuroBERT61...	0.697	0.561	0.621	0.803	0.647	0.717
DT4H_NL	EuroBERT61...	0.699	0.558	0.621	0.802	0.641	0.713
SuSh	run	0.329	0.409	0.365	0.436	0.532	0.479
blue	run	0.290	0.259	0.274	0.739	0.659	0.696
nlp-fbk	run	0.000	0.000	0.000	0.602	0.559	0.579
HSE NLP	run	0.000	0.000	0.000	0.079	0.000	0.000
Language: Spanish (es)							
Entity: Disease							
BIT.UA	ensemble-a...	0.810	0.839	0.824	0.866	0.897	0.881
BIT.UA	ensemble-n...	0.796	0.842	0.818	0.856	0.905	0.880
LSI_UNED	es_conf2	0.785	0.843	0.813	0.850	0.897	0.873
BIT.UA	roberta-ba...	0.785	0.839	0.811	0.847	0.907	0.876
LSI_UNED	es_conf1	0.792	0.823	0.807	0.855	0.876	0.865
SIEMENS	es_disease...	0.817	0.797	0.807	0.871	0.850	0.861
BIT.UA	ensemble-x...	0.800	0.811	0.805	0.861	0.875	0.868
BIT.UA	xlm-robert...	0.775	0.818	0.796	0.843	0.892	0.867
SIEMENS	es_disease...	0.752	0.825	0.786	0.821	0.899	0.859
SIEMENS	es_disease...	0.763	0.767	0.765	0.836	0.842	0.839
SIEMENS	es_disease...	0.754	0.769	0.761	0.831	0.850	0.840
Parallia	run	0.719	0.787	0.751	0.807	0.882	0.843
SINAI	run	0.746	0.744	0.745	0.838	0.828	0.833
SIEMENS	es_disease...	0.800	0.697	0.745	0.884	0.769	0.822
Discovery@FI	run	0.695	0.763	0.728	0.796	0.867	0.830
Enigma	disease_cl...	0.628	0.752	0.684	0.767	0.878	0.819
Enigma	rigoberta-...	0.630	0.746	0.683	0.771	0.872	0.819
DT4H_NL	EuroBERT61...	0.710	0.617	0.660	0.805	0.702	0.750
DT4H_NL	EuroBERT61...	0.713	0.603	0.653	0.818	0.694	0.751
DT4H_NL	DeBERTa_mu...	0.664	0.605	0.633	0.765	0.700	0.731
IIC-UC3M	run	0.407	0.382	0.394	0.755	0.707	0.730
blue	run	0.351	0.343	0.347	0.807	0.785	0.795
nlp-fbk	run	0.002	0.002	0.002	0.661	0.635	0.648
Entity: Procedure							

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
BIT.UA	ensemble-a...	0.814	0.813	0.813	0.880	0.880	0.880
BIT.UA	ensemble-n...	0.801	0.815	0.808	0.872	0.886	0.879
LSI_UNED	conf2	0.802	0.805	0.803	0.873	0.866	0.870
LSI_UNED	conf1	0.801	0.800	0.801	0.870	0.856	0.863
BIT.UA	roberta-ba...	0.790	0.809	0.800	0.866	0.889	0.877
BIT.UA	ensemble-x...	0.782	0.804	0.793	0.857	0.881	0.869
SIEMENS	es_procedu...	0.824	0.763	0.792	0.890	0.823	0.855
BIT.UA	xlm-robert...	0.768	0.799	0.783	0.847	0.884	0.865
SIEMENS	es_procedu...	0.765	0.797	0.781	0.849	0.882	0.865
SIEMENS	es_procedu...	0.785	0.763	0.774	0.858	0.832	0.845
SIEMENS	es_procedu...	0.769	0.768	0.769	0.844	0.843	0.843
SIEMENS	es_procedu...	0.782	0.756	0.769	0.867	0.833	0.850
Parallia	run	0.719	0.792	0.754	0.811	0.888	0.848
Discovery@FI	run	0.675	0.766	0.717	0.785	0.879	0.829
SINAI	run	0.746	0.679	0.711	0.844	0.765	0.803
Enigma	rigoberta-...	0.620	0.744	0.676	0.762	0.873	0.814
Enigma	procedure_...	0.614	0.745	0.673	0.758	0.876	0.813
DT4H_NL	EuroBERT61...	0.747	0.599	0.665	0.845	0.679	0.753
DT4H_NL	EuroBERT61...	0.727	0.602	0.658	0.827	0.688	0.751
IIC-UC3M	run	0.500	0.488	0.494	0.772	0.749	0.761
blue	run	0.435	0.460	0.447	0.789	0.828	0.808
nlp-fbk	run	0.000	0.000	0.000	0.698	0.686	0.692
Entity: Symptom							
BIT.UA	ensemble-2...	0.784	0.774	0.779	0.848	0.839	0.843
BIT.UA	ensemble-2...	0.764	0.779	0.771	0.835	0.851	0.843
BIT.UA	ensemble-7...	0.760	0.764	0.762	0.828	0.835	0.831
BIT.UA	roberta-ba...	0.750	0.771	0.760	0.826	0.853	0.839
LSI_UNED	es_conf2	0.747	0.767	0.757	0.823	0.834	0.828
LSI_UNED	es_conf1	0.763	0.749	0.756	0.837	0.807	0.821
BIT.UA	xlm-robert...	0.735	0.772	0.753	0.812	0.857	0.834
SIEMENS	es_symptom...	0.770	0.725	0.747	0.836	0.787	0.811
SIEMENS	es_symptom...	0.731	0.735	0.733	0.811	0.816	0.814
SIEMENS	es_symptom...	0.727	0.715	0.721	0.820	0.805	0.812
SIEMENS	es_symptom...	0.701	0.711	0.706	0.790	0.802	0.795
SIEMENS	es_symptom...	0.699	0.710	0.705	0.784	0.797	0.790
Parallia	run	0.627	0.718	0.670	0.744	0.846	0.791
SINAI	run	0.685	0.637	0.660	0.780	0.727	0.752
Discovery@FI	run	0.578	0.677	0.624	0.718	0.825	0.767
DT4H_NL	EuroBERT61...	0.668	0.560	0.609	0.770	0.648	0.704
DT4H_NL	EuroBERT61...	0.661	0.563	0.609	0.766	0.654	0.706
Enigma	rigoberta-...	0.475	0.640	0.545	0.657	0.813	0.727
Enigma	symptom_cl...	0.476	0.638	0.545	0.657	0.810	0.726
IIC-UC3M	run	0.385	0.331	0.356	0.725	0.623	0.670
blue	run	0.313	0.314	0.314	0.727	0.728	0.728
nlp-fbk	run	0.000	0.000	0.000	0.577	0.519	0.546
Language: Italian (it)							

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Entity: Disease							
BIT.UA	ensemble-x...	0.803	0.702	0.749	0.884	0.777	0.827
BIT.UA	xlm-robert...	0.759	0.730	0.744	0.856	0.828	0.842
BIT.UA	ensemble-a...	0.790	0.692	0.738	0.874	0.767	0.817
BIT.UA	ensemble-n...	0.746	0.707	0.726	0.846	0.799	0.822
LSI_UNED	conf1_biob...	0.706	0.725	0.715	0.820	0.819	0.820
LSI_UNED	conf1_medp...	0.716	0.710	0.713	0.825	0.801	0.813
SIEMENS	it_disease...	0.725	0.693	0.709	0.821	0.787	0.803
Enigma	ensemble_i...	0.738	0.676	0.706	0.840	0.754	0.795
Parallia	run	0.683	0.723	0.703	0.802	0.846	0.824
SIEMENS	it_disease...	0.706	0.699	0.702	0.813	0.805	0.809
BIT.UA	Italian-Bi...	0.720	0.683	0.701	0.828	0.786	0.806
SIEMENS	it_disease...	0.705	0.693	0.699	0.826	0.814	0.820
SIEMENS	it_disease...	0.684	0.701	0.692	0.804	0.826	0.815
SIEMENS	it_disease...	0.697	0.659	0.677	0.811	0.767	0.788
Discovery@FI	run	0.640	0.700	0.669	0.769	0.832	0.799
LemonHeard	run2	0.643	0.664	0.653	0.793	0.815	0.804
LemonHeard	run1	0.538	0.672	0.597	0.696	0.856	0.767
DT4H_NL	EuroBERT61...	0.631	0.552	0.589	0.791	0.694	0.739
DT4H_NL	EuroBERT61...	0.630	0.527	0.574	0.805	0.677	0.736
DT4H_NL	DeBERTa_mu...	0.596	0.541	0.567	0.754	0.688	0.719
Enigma	team_enigm...	0.369	0.330	0.348	0.822	0.735	0.776
blue	run	0.184	0.186	0.185	0.624	0.629	0.626
nlp-fbk	run	0.001	0.001	0.001	0.659	0.564	0.608
Entity: Procedure							
BIT.UA	ensemble-x...	0.749	0.728	0.739	0.857	0.834	0.845
BIT.UA	xlm-robert...	0.734	0.719	0.726	0.847	0.836	0.842
BIT.UA	ensemble-a...	0.761	0.691	0.724	0.865	0.785	0.823
BIT.UA	ensemble-n...	0.734	0.685	0.709	0.849	0.790	0.819
Parallia	run	0.674	0.722	0.697	0.802	0.855	0.828
SIEMENS	it_procedu...	0.692	0.690	0.691	0.815	0.808	0.811
SIEMENS	it_procedu...	0.711	0.670	0.690	0.833	0.781	0.806
BIT.UA	Italian-Bi...	0.698	0.671	0.684	0.827	0.797	0.811
LSI_UNED	conf1_medp...	0.712	0.632	0.670	0.850	0.744	0.793
Discovery@FI	run	0.628	0.705	0.664	0.765	0.846	0.804
LSI_UNED	conf1_biob...	0.689	0.638	0.662	0.843	0.764	0.801
SIEMENS	it_procedu...	0.649	0.633	0.641	0.820	0.800	0.810
SIEMENS	it_procedu...	0.629	0.640	0.634	0.786	0.794	0.790
SIEMENS	it_procedu...	0.612	0.594	0.603	0.789	0.758	0.773
LemonHeard	run2	0.545	0.572	0.558	0.770	0.799	0.784
LemonHeard	run1	0.462	0.585	0.516	0.685	0.844	0.756
DT4H_NL	EuroBERT61...	0.570	0.472	0.516	0.801	0.667	0.728
DT4H_NL	EuroBERT61...	0.576	0.465	0.514	0.819	0.662	0.733
Enigma	ensemble_i...	0.359	0.420	0.387	0.429	0.506	0.465
Enigma	ensemble_i...	0.372	0.373	0.372	0.415	0.416	0.416
blue	run	0.292	0.234	0.260	0.783	0.637	0.702
nlp-fbk	run	0.001	0.000	0.000	0.629	0.533	0.577

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
Entity: Symptom							
BIT.UA	ensemble-x...	0.734	0.648	0.689	0.842	0.747	0.792
BIT.UA	xlm-robert...	0.704	0.664	0.683	0.825	0.782	0.803
BIT.UA	ensemble-a...	0.742	0.615	0.673	0.846	0.704	0.768
BIT.UA	ensemble-n...	0.692	0.615	0.651	0.812	0.722	0.764
LSI_UNED	conf1_biob...	0.690	0.617	0.651	0.817	0.708	0.758
LSI_UNED	conf1_medp...	0.680	0.624	0.651	0.812	0.719	0.762
SIEMENS	it_symptom...	0.643	0.624	0.633	0.782	0.757	0.769
SIEMENS	it_symptom...	0.649	0.616	0.632	0.781	0.742	0.761
BIT.UA	Italian-Bi...	0.654	0.595	0.623	0.790	0.722	0.754
SIEMENS	it_symptom...	0.638	0.596	0.616	0.775	0.724	0.748
Enigma	run	0.665	0.569	0.613	0.796	0.680	0.733
SIEMENS	it_symptom...	0.636	0.578	0.606	0.772	0.699	0.734
SIEMENS	it_symptom...	0.596	0.604	0.600	0.749	0.754	0.751
Parallia	run	0.565	0.634	0.598	0.725	0.806	0.764
LemonHeard	run2	0.593	0.576	0.584	0.750	0.724	0.737
Discovery@FI	run	0.521	0.607	0.561	0.696	0.791	0.741
DT4H_NL	EuroBERT61...	0.601	0.490	0.540	0.764	0.625	0.688
DT4H_NL	EuroBERT61...	0.609	0.479	0.536	0.773	0.612	0.683
LemonHeard	run1	0.447	0.589	0.509	0.622	0.796	0.698
blue	run	0.183	0.135	0.156	0.662	0.496	0.567
nlp-fbk	run	0.000	0.000	0.000	0.544	0.510	0.526
Language: Dutch (nl)							
Entity: Disease							
BIT.UA	ensemble-x...	0.736	0.739	0.737	0.820	0.827	0.823
BIT.UA	xlm-robert...	0.728	0.733	0.731	0.816	0.826	0.821
Enigma	disease-5	0.765	0.665	0.712	0.843	0.732	0.784
BIT.UA	ensemble-a...	0.739	0.683	0.710	0.822	0.764	0.792
SIEMENS	nl_disease...	0.708	0.687	0.697	0.799	0.780	0.789
SIEMENS	nl_disease...	0.696	0.690	0.693	0.794	0.790	0.792
Enigma	team_enigm...	0.723	0.654	0.687	0.811	0.732	0.769
BIT.UA	ensemble-x...	0.697	0.674	0.685	0.793	0.768	0.781
Parallia	run	0.657	0.700	0.678	0.767	0.820	0.793
SIEMENS	nl_disease...	0.709	0.645	0.676	0.799	0.731	0.764
SIEMENS	nl_disease...	0.691	0.657	0.674	0.787	0.751	0.769
BIT.UA	MedRoBERTa...	0.661	0.667	0.664	0.766	0.778	0.772
Discovery@FI	run	0.639	0.684	0.661	0.752	0.800	0.775
DT4H_NL	robbert202...	0.638	0.683	0.660	0.737	0.791	0.763
Dr-BERT-NL	run	0.664	0.651	0.657	0.768	0.756	0.762
DT4H_NL	robbert202...	0.648	0.663	0.655	0.749	0.772	0.760
SIEMENS	nl_disease...	0.637	0.661	0.649	0.744	0.772	0.758
DT4H_NL	DeBERTa_mu...	0.600	0.676	0.635	0.694	0.788	0.738
Lotus Orchid	all_xlmrb_...	0.622	0.626	0.624	0.736	0.739	0.737
Lotus Orchid	nl_xlmrb_d...	0.612	0.615	0.614	0.728	0.730	0.729
Lotus Orchid	nl_medrobe...	0.583	0.578	0.580	0.716	0.715	0.715
DT4H_NL	EuroBERT61...	0.633	0.524	0.573	0.743	0.622	0.677

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
DT4H_NL	EuroBERT61...	0.621	0.532	0.573	0.726	0.628	0.673
Lotus Orchid	nl+synth_m...	0.535	0.564	0.549	0.671	0.706	0.688
Lotus Orchid	nl_robbert...	0.533	0.549	0.541	0.665	0.690	0.677
Enigma	disease-4	0.511	0.558	0.534	0.637	0.697	0.666
Enigma	disease-2	0.444	0.336	0.383	0.654	0.501	0.568
blue	run	0.378	0.359	0.368	0.752	0.716	0.734
Enigma	disease-3	0.591	0.157	0.248	0.719	0.192	0.303
nlp-fbk	run	0.000	0.000	0.000	0.617	0.498	0.551
Entity: Procedure							
BIT.UA	ensemble-x...	0.723	0.714	0.718	0.829	0.822	0.825
BIT.UA	xlm-robert...	0.708	0.710	0.709	0.817	0.827	0.822
Enigma	procedure-...	0.762	0.662	0.708	0.854	0.740	0.793
BIT.UA	ensemble-a...	0.750	0.658	0.701	0.845	0.746	0.792
SIEMENS	nl_procedu...	0.720	0.677	0.698	0.825	0.777	0.800
SIEMENS	nl_procedu...	0.704	0.683	0.693	0.814	0.792	0.803
Parallia	run	0.678	0.707	0.692	0.794	0.829	0.811
SIEMENS	nl_procedu...	0.763	0.628	0.689	0.850	0.702	0.769
Enigma	team_enigm...	0.722	0.653	0.685	0.826	0.744	0.783
SIEMENS	nl_procedu...	0.722	0.645	0.681	0.822	0.736	0.777
BIT.UA	ensemble-n...	0.709	0.647	0.677	0.818	0.749	0.782
Discovery@FI	run	0.648	0.700	0.673	0.772	0.831	0.801
Dr-BERT-NL	run	0.685	0.659	0.672	0.797	0.770	0.783
BIT.UA	MedRoBERTa...	0.670	0.646	0.657	0.791	0.770	0.781
DT4H_NL	robbert202...	0.657	0.654	0.655	0.775	0.778	0.776
SIEMENS	nl_procedu...	0.673	0.635	0.653	0.794	0.753	0.773
Lotus Orchid	all_xlmrb_...	0.656	0.593	0.623	0.792	0.715	0.752
Lotus Orchid	nl_xlmrb_p...	0.627	0.590	0.608	0.767	0.724	0.745
DT4H_NL	EuroBERT61...	0.694	0.523	0.596	0.810	0.616	0.700
Lotus Orchid	nl_medrobe...	0.604	0.576	0.590	0.750	0.720	0.735
DT4H_NL	EuroBERT61...	0.693	0.509	0.587	0.813	0.602	0.692
Lotus Orchid	nl_robbert...	0.567	0.587	0.577	0.696	0.728	0.711
Enigma	procedure-...	0.578	0.573	0.576	0.694	0.692	0.693
Lotus Orchid	nl+synth_m...	0.534	0.566	0.549	0.674	0.716	0.694
blue	run	0.415	0.400	0.407	0.735	0.712	0.723
Enigma	procedure-...	0.439	0.377	0.405	0.665	0.567	0.612
Enigma	procedure-...	0.602	0.183	0.281	0.685	0.212	0.324
nlp-fbk	run	0.001	0.000	0.000	0.688	0.554	0.614
Entity: Symptom							
BIT.UA	ensemble-8...	0.663	0.633	0.647	0.783	0.750	0.766
BIT.UA	xlm-robert...	0.642	0.624	0.633	0.770	0.751	0.760
BIT.UA	ensemble-2...	0.675	0.582	0.625	0.787	0.680	0.730
Enigma	symptom-5	0.695	0.536	0.606	0.809	0.622	0.703
SIEMENS	nl_symptom...	0.622	0.586	0.603	0.745	0.702	0.723
SIEMENS	nl_symptom...	0.613	0.584	0.598	0.743	0.708	0.725
BIT.UA	ensemble-1...	0.624	0.569	0.595	0.753	0.685	0.717
SIEMENS	nl_symptom...	0.651	0.540	0.590	0.771	0.639	0.699
Enigma	team_enigm...	0.649	0.528	0.582	0.780	0.629	0.697

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
DT4H_NL	robbert202...	0.562	0.588	0.575	0.689	0.721	0.705
BIT.UA	MedRoBERTa...	0.579	0.565	0.572	0.719	0.705	0.712
SIEMENS	nl_symptom...	0.624	0.523	0.569	0.761	0.638	0.694
Dr-BERT-NL	run	0.580	0.551	0.565	0.728	0.690	0.708
Parallia	run	0.542	0.587	0.564	0.699	0.754	0.726
Discovery@FI	run	0.522	0.582	0.550	0.680	0.743	0.710
SIEMENS	nl_symptom...	0.557	0.537	0.547	0.712	0.684	0.698
Lotus Orchid	all_xlmrb_...	0.551	0.479	0.513	0.710	0.616	0.660
DT4H_NL	EuroBERT61...	0.608	0.425	0.500	0.751	0.527	0.619
DT4H_NL	EuroBERT61...	0.601	0.422	0.496	0.748	0.529	0.620
Lotus Orchid	nl_xlmrb_s...	0.548	0.441	0.489	0.712	0.572	0.635
Lotus Orchid	nl_medrobe...	0.552	0.430	0.483	0.727	0.570	0.639
Enigma	symptom-4	0.469	0.495	0.481	0.621	0.651	0.636
Lotus Orchid	nl_robbert...	0.502	0.417	0.456	0.672	0.562	0.612
Lotus Orchid	nl+synth_m...	0.439	0.403	0.420	0.616	0.560	0.587
Enigma	symptom-2	0.338	0.256	0.291	0.570	0.426	0.488
Enigma	symptom-3	0.622	0.186	0.286	0.741	0.223	0.343
blue	run	0.295	0.256	0.274	0.688	0.597	0.639
nlp-fbk	run	0.000	0.000	0.000	0.418	0.224	0.291
Language: Romanian (ro)							
Entity: Disease							
BIT.UA	ensemble-x...	0.779	0.762	0.770	0.868	0.855	0.862
BIT.UA	xlm-robert...	0.772	0.754	0.763	0.866	0.852	0.859
SIEMENS	ro_disease...	0.729	0.731	0.730	0.823	0.830	0.826
SIEMENS	ro_disease...	0.742	0.717	0.729	0.835	0.811	0.823
Parallia	run	0.696	0.743	0.719	0.804	0.859	0.831
BIT.UA	ensemble-a...	0.753	0.685	0.718	0.845	0.772	0.807
SIEMENS	ro_disease...	0.715	0.710	0.713	0.827	0.822	0.825
SIEMENS	ro_disease...	0.740	0.683	0.711	0.830	0.771	0.799
SIEMENS	ro_disease...	0.719	0.701	0.710	0.816	0.798	0.807
BIT.UA	ensemble-n...	0.723	0.675	0.698	0.825	0.772	0.797
Enigma	disease	0.685	0.699	0.692	0.809	0.827	0.818
Enigma	ensemble-r...	0.682	0.699	0.690	0.805	0.828	0.816
SIEMENS	ro_disease...	0.691	0.684	0.687	0.809	0.803	0.806
Discovery@FI	run	0.658	0.716	0.686	0.781	0.844	0.811
BIT.UA	bert-base-...	0.691	0.667	0.679	0.800	0.777	0.788
DT4H_NL	DeBERTa_mu...	0.639	0.627	0.633	0.767	0.757	0.762
DT4H_NL	EuroBERT61...	0.679	0.566	0.617	0.812	0.684	0.743
DT4H_NL	EuroBERT61...	0.679	0.534	0.598	0.823	0.654	0.729
blue	run	0.199	0.194	0.196	0.666	0.650	0.658
Enigma	gliner-ro-...	0.824	0.057	0.107	0.872	0.061	0.114
nlp-fbk	run	0.001	0.001	0.001	0.647	0.624	0.635
Entity: Procedure							
BIT.UA	ensemble-x...	0.765	0.740	0.752	0.866	0.840	0.853
BIT.UA	xlm-robert...	0.751	0.732	0.741	0.858	0.841	0.849
BIT.UA	ensemble-a...	0.782	0.691	0.734	0.872	0.771	0.818

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
SIEMENS	ro_procedu...	0.751	0.714	0.732	0.855	0.811	0.832
SIEMENS	ro_procedu...	0.736	0.723	0.729	0.848	0.830	0.839
Parallia	run	0.706	0.739	0.722	0.822	0.858	0.840
BIT.UA	ensemble-n...	0.751	0.677	0.712	0.855	0.769	0.809
SIEMENS	ro_procedu...	0.743	0.679	0.710	0.848	0.774	0.810
Enigma	procedure	0.700	0.718	0.709	0.823	0.842	0.833
SIEMENS	ro_procedu...	0.708	0.686	0.697	0.832	0.806	0.818
BIT.UA	bert-base-...	0.724	0.668	0.695	0.839	0.777	0.807
SIEMENS	ro_procedu...	0.729	0.659	0.692	0.850	0.764	0.804
Discovery@FI	run	0.664	0.720	0.691	0.796	0.852	0.823
SIEMENS	ro_procedu...	0.703	0.678	0.691	0.831	0.801	0.816
Enigma	xlmr-ro-pr...	0.658	0.722	0.689	0.792	0.856	0.823
DT4H_NL	EuroBERT61...	0.721	0.568	0.636	0.851	0.672	0.751
DT4H_NL	EuroBERT61...	0.713	0.569	0.633	0.840	0.675	0.748
blue	run	0.219	0.191	0.204	0.679	0.600	0.637
nlp-fbk	run	0.001	0.000	0.001	0.637	0.497	0.558
Entity: Symptom							
BIT.UA	ensemble-x...	0.768	0.650	0.704	0.858	0.728	0.787
BIT.UA	xlm-robert...	0.720	0.682	0.701	0.830	0.788	0.809
SIEMENS	ro_symptom...	0.699	0.649	0.673	0.804	0.746	0.774
SIEMENS	ro_symptom...	0.697	0.648	0.671	0.804	0.748	0.775
BIT.UA	ensemble-a...	0.765	0.597	0.671	0.854	0.667	0.749
BIT.UA	ensemble-n...	0.715	0.604	0.655	0.823	0.693	0.753
SIEMENS	ro_symptom...	0.672	0.635	0.653	0.792	0.749	0.770
SIEMENS	ro_symptom...	0.722	0.566	0.634	0.829	0.650	0.729
Enigma	run	0.625	0.635	0.630	0.762	0.773	0.768
SIEMENS	ro_symptom...	0.711	0.561	0.627	0.824	0.649	0.726
Parallia	run	0.604	0.647	0.625	0.750	0.800	0.774
BIT.UA	bert-base-...	0.659	0.591	0.623	0.784	0.704	0.742
SIEMENS	ro_symptom...	0.621	0.621	0.621	0.750	0.744	0.747
Discovery@FI	run	0.562	0.627	0.593	0.722	0.786	0.753
DT4H_NL	EuroBERT61...	0.651	0.496	0.563	0.785	0.601	0.681
DT4H_NL	EuroBERT61...	0.647	0.482	0.552	0.784	0.586	0.671
blue	run	0.285	0.211	0.243	0.755	0.568	0.648
nlp-fbk	run	0.001	0.000	0.000	0.415	0.121	0.188
Language: Swedish (sv)							
Entity: Disease							
BIT.UA	ensemble-x...	0.759	0.717	0.738	0.850	0.808	0.829
BIT.UA	xlm-robert...	0.736	0.726	0.731	0.835	0.828	0.831
BIT.UA	ensemble-a...	0.756	0.686	0.719	0.850	0.773	0.809
BIT.UA	ensemble-n...	0.725	0.666	0.694	0.828	0.761	0.793
SIEMENS	sv_disease...	0.699	0.686	0.692	0.802	0.791	0.796
SIEMENS	sv_disease...	0.687	0.694	0.690	0.795	0.805	0.800
SIEMENS	sv_disease...	0.687	0.687	0.687	0.796	0.800	0.798
Parallia	run	0.661	0.702	0.681	0.783	0.830	0.806
BIT.UA	bert-base-...	0.694	0.659	0.676	0.807	0.769	0.788

Continued on next page

Team	Run	Strict			Relaxed (character)		
		P	R	F1	P	R	F1
SIEMENS	sv_disease...	0.687	0.662	0.674	0.796	0.769	0.782
Discovery@FI	run	0.640	0.685	0.662	0.767	0.812	0.789
SIEMENS	sv_disease...	0.691	0.617	0.652	0.814	0.729	0.769
DT4H_NL	EuroBERT61...	0.652	0.536	0.588	0.787	0.649	0.712
DT4H_NL	DeBERTa_mu...	0.605	0.561	0.583	0.738	0.685	0.710
DT4H_NL	EuroBERT61...	0.658	0.515	0.578	0.793	0.627	0.700
blue	run	0.199	0.197	0.198	0.606	0.597	0.602
nlp-fbk	run	0.000	0.000	0.000	0.551	0.578	0.564
Entity: Procedure							
BIT.UA	ensemble-x...	0.744	0.732	0.738	0.848	0.838	0.843
BIT.UA	ensemble-a...	0.766	0.697	0.730	0.861	0.785	0.821
BIT.UA	xlm-robert...	0.726	0.726	0.726	0.837	0.843	0.840
SIEMENS	sv_procedu...	0.709	0.701	0.705	0.826	0.815	0.820
BIT.UA	ensemble-n...	0.725	0.686	0.705	0.832	0.790	0.810
SIEMENS	sv_procedu...	0.715	0.694	0.704	0.830	0.807	0.819
SIEMENS	sv_procedu...	0.705	0.699	0.702	0.821	0.817	0.819
Parallia	run	0.681	0.723	0.702	0.804	0.853	0.828
SIEMENS	sv_procedu...	0.710	0.671	0.690	0.825	0.780	0.802
BIT.UA	bert-base-...	0.698	0.681	0.690	0.817	0.801	0.809
SIEMENS	sv_procedu...	0.705	0.671	0.687	0.826	0.783	0.804
Discovery@FI	run	0.658	0.718	0.687	0.783	0.849	0.815
DT4H_NL	EuroBERT61...	0.701	0.556	0.620	0.831	0.664	0.738
DT4H_NL	EuroBERT61...	0.698	0.557	0.620	0.832	0.669	0.741
blue	run	0.286	0.252	0.268	0.691	0.616	0.652
nlp-fbk	run	0.000	0.000	0.000	0.655	0.422	0.513
Entity: Symptom							
BIT.UA	ensemble-x...	0.717	0.669	0.692	0.826	0.771	0.797
BIT.UA	xlm-robert...	0.689	0.677	0.683	0.806	0.793	0.800
BIT.UA	ensemble-a...	0.739	0.606	0.666	0.839	0.686	0.755
SIEMENS	sv_symptom...	0.664	0.653	0.658	0.782	0.766	0.774
SIEMENS	sv_symptom...	0.659	0.646	0.653	0.788	0.769	0.778
SIEMENS	sv_symptom...	0.663	0.624	0.643	0.779	0.732	0.755
BIT.UA	ensemble-n...	0.686	0.601	0.641	0.809	0.707	0.755
Parallia	run	0.590	0.651	0.619	0.734	0.802	0.767
BIT.UA	bert-base-...	0.640	0.598	0.618	0.775	0.724	0.749
SIEMENS	sv_symptom...	0.626	0.608	0.617	0.763	0.735	0.749
SIEMENS	sv_symptom...	0.682	0.560	0.615	0.803	0.656	0.722
Discovery@FI	run	0.562	0.638	0.598	0.712	0.788	0.748
DT4H_NL	EuroBERT61...	0.641	0.467	0.540	0.793	0.578	0.668
DT4H_NL	EuroBERT61...	0.642	0.451	0.530	0.791	0.557	0.654
blue	run	0.324	0.243	0.278	0.765	0.574	0.656
nlp-fbk	run	0.000	0.000	0.000	0.442	0.223	0.296

Table 8: Overall MultiClinNER results by language and entity. *Note.* Results are reported for the overall corpus. Rows in italics indicate runs for which both Strict F1 and Character F1 are 0.000, which usually correspond to submission errors. Run names longer than 10 characters are abbreviated with ellipses.

B.2 MultiClinCorpus

Entity	Team	Run	P	R	F1	Team	Run	P	R	F1
		Czech					**English**			
Disease	Parallia	ClinicalAligner26AM-S3	0.845	0.856	0.851	Parallia	ClinicalAligner26AM-S3	0.891	0.901	0.896
	Parallia	ClinicalAligner26AM-S2	0.819	0.830	0.824	Parallia	ClinicalAligner26AM-S2	0.880	0.889	0.884
	blue	cognate-fuzzy-v1	0.103	0.022	0.036	ICB-UMA	run4	0.738	0.735	0.737
						ICB-UMA	run3	0.738	0.733	0.736
						ICB-UMA	run2	0.689	0.689	0.689
						ICB-UMA	run1	0.415	0.411	0.413
						blue	cognate-fuzzy-v1	0.247	0.098	0.140
Procedure	Parallia	ClinicalAligner26AM-S3	0.817	0.820	0.819	Parallia	ClinicalAligner26AM-S3	0.835	0.847	0.841
	Parallia	ClinicalAligner26AM-S2	0.805	0.807	0.806	Parallia	ClinicalAligner26AM-S2	0.835	0.846	0.840
	blue	cognate-fuzzy-v1	0.119	0.026	0.043	ICB-UMA	run	0.552	0.545	0.549
						blue	cognate-fuzzy-v1	0.158	0.052	0.079
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.876	0.882	0.879
	Parallia	ClinicalAligner26AM-S2	0.790	0.791	0.791	Parallia	ClinicalAligner26AM-S2	0.862	0.867	0.864
	blue	cognate-fuzzy-v1	0.085	0.015	0.025	ICB-UMA	run	0.540	0.537	0.538
						blue	cognate-fuzzy-v1	0.238	0.074	0.113
		Swedish					**Italian**			
Disease	Parallia	ClinicalAligner26AM-S2	0.822	0.834	0.828	Parallia	ClinicalAligner26AM-S3	0.876	0.887	0.882
	Parallia	ClinicalAligner26AM-S3	0.819	0.831	0.825	Parallia	ClinicalAligner26AM-S2	0.838	0.847	0.843
	blue	cognate-fuzzy-v1	0.227	0.065	0.101	blue	cognate-fuzzy-v1	0.242	0.193	0.215
Procedure	Parallia	ClinicalAligner26AM-S3	0.803	0.809	0.806	Parallia	ClinicalAligner26AM-S3	0.792	0.806	0.799
	Parallia	ClinicalAligner26AM-S2	0.792	0.798	0.795	Parallia	ClinicalAligner26AM-S2	0.764	0.778	0.771
	blue	cognate-fuzzy-v1	0.164	0.038	0.062	blue	cognate-fuzzy-v1	0.246	0.172	0.203
Symptom	Parallia	ClinicalAligner26AM-S3	0.810	0.812	0.811	Parallia	ClinicalAligner26AM-S3	0.829	0.833	0.831
	Parallia	ClinicalAligner26AM-S2	0.798	0.799	0.799	Parallia	ClinicalAligner26AM-S2	0.785	0.789	0.787
	blue	cognate-fuzzy-v1	0.248	0.054	0.089	blue	cognate-fuzzy-v1	0.255	0.179	0.211
		Dutch					**Romanian**			
Disease	Parallia	ClinicalAligner26AM-S3	0.818	0.822	0.820	Parallia	ClinicalAligner26AM-S3	0.894	0.902	0.898
	Parallia	ClinicalAligner26AM-S2	0.802	0.806	0.804	Parallia	ClinicalAligner26AM-S2	0.875	0.882	0.878
	blue	cognate-fuzzy-v1	0.146	0.052	0.077	blue	cognate-fuzzy-v1	0.151	0.105	0.124
Procedure	Parallia	ClinicalAligner26AM-S3	0.774	0.766	0.770	Parallia	ClinicalAligner26AM-S3	0.850	0.861	0.856
	Parallia	ClinicalAligner26AM-S2	0.759	0.751	0.755	Parallia	ClinicalAligner26AM-S2	0.844	0.854	0.849
	blue	cognate-fuzzy-v1	0.096	0.029	0.044	blue	cognate-fuzzy-v1	0.156	0.101	0.122
Symptom	Parallia	ClinicalAligner26AM-S3	0.767	0.769	0.768	Parallia	ClinicalAligner26AM-S3	0.858	0.864	0.861
	Parallia	ClinicalAligner26AM-S2	0.762	0.763	0.763	Parallia	ClinicalAligner26AM-S2	0.831	0.836	0.834
	blue	cognate-fuzzy-v1	0.091	0.022	0.036	blue	cognate-fuzzy-v1	0.136	0.076	0.098

Table 9: Performance of all participating systems for the MultiClinCorpus subtask under strict evaluation, reported for each entity type and language. Precision (P), Recall (R), and F1-score (F1) are computed using exact span and label matching after annotation projection.