Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 283-296 recibido 06-04-2025 revisado 10-05-2025 aceptado 19-05-2025

EriBERTa Private Surpasses her Public Alter Ego: Enhancing a Bilingual Pretrained Encoder with Limited Private Medical Data

EriBERTa Privada Supera su Alter Ego Pública: Mejorando un Codificador Bilingüe Preentrenado con Datos Médicos Privados Limitados

Iker De la Iglesia, Adrián Sánchez-Freire, Oier Urquijo-Durán, Ander Barrena, Aitziber Atutxa

HiTZ Basque Center for Language Technology, University of the Basque Country UPV/EHU

{iker.delaiglesia, ander.barrena, aitziber.atutxa}@ehu.eus

Abstract: The secondary use of clinical reports is essential for improving patient care. While NLP tools have become instrumental in extracting insights from such reports, domain-specific language models for clinical Spanish remain scarce. Therefore, we introduce EriBERTa, the first open-source bilingual clinical language model for English and Spanish, designed to advance clinical NLP in under-resourced settings. We evaluate its performance across multiple dimensions: public vs. proprietary pretraining data, data availability, and cross-lingual transfer. Results show that pretraining on in-domain Electronic Health Records yields strong gains, especially for complex tasks like clinical document section identification. EriBERTa also performs well on monolingual tasks and transfers effectively across languages, making it a valuable tool for multilingual clinical NLP. The model is publicly released to support further research.
Keywords: Evidence-Based Medicine, Language Models, Cross-Lingual Transfer.

Resumen: El uso secundario de los informes clínicos es esencial para mejorar la atención al paciente. Si bien las herramientas de PLN se han vuelto fundamentales para extraer información de dichos informes, los Modelos del Lenguaje específicos de dominio para el español clínico siguen siendo escasos. Presentamos EriBERTa, el primer Modelo del Lenguaje clínico bilingüe de código abierto para inglés y español, diseñado para impulsar el Procesamiento del Lenguaje Clínico en entornos de bajos recursos. Evaluamos su rendimiento en múltiples dimensiones: datos de preentrenamiento públicos y privados, disponibilidad de datos y transferencia interlingüística. Los resultados muestran que el preentrenamiento en Informes Clínicos Electrónicos dentro del dominio produce importantes mejoras, especialmente en tareas complejas como la identificación de secciones en informes clínicos. EriBERTa también muestra buen rendimiento en tareas monolingües y transfiere el conocimiento adquirido eficazmente entre idiomas, lo que lo convierte en una herramienta valiosa para el PLN clínico multilingüe. El modelo se publica para apoyar futuras investigaciones.
Palabras clave: Medicina basada en evidencia, modelos del lenguaje, transferencia interlingüística.

1 Introduction

Processing relevant information from clinical reports in healthcare systems presents particular challenges given the domain-specific language, extensive use of abbreviations, and their unstructured nature. **Natural Language Processing (NLP)** techniques have emerged as powerful tools for tackling these challenges and unlocking the potential of clinical data.

ISSN 1135-5948 DOI 10.26342/2025-75-21

Furthermore, as highlighted by Goenaga et al. (2021), the advancement of technology in automatic processing has occurred simultaneously with the exponential increase in digitized data by healthcare systems. Official reports substantiate this trend. For example, in 2016, the proportion of primary care consultations using electronic health records was approximately 80% on average in 15 European Union (EU) countries (EUH, 2020), a figure that surged to 96% in the United States (U.S.) by 2020 (EUH, 2020; ITD, 2021). This digitization of healthcare systems plays a pivotal role in enhancing clinical and translational studies, making interoperability and information exchange between systems more crucial than ever. Consequently, public policies and recommendations advocate for the exchange of health information (ONC, 2015; Hor, 2019; ONC, 2020). Numerous standards, including *open*EHR (ope, 2020), HL7-FHIR (HL7, 2019b), HL7 CDA/CCR (HL7, 2019a), *International Classification of Diseases* (ICD-10, ICD-11 (ICD, 2022)), and *Concept Unique Identifier* (CUIs (I.H.T.S.D.O, 2022)), exemplify the ongoing standardization efforts.

However, as highlighted by López-Úbeda (2022), standardization presents several challenges due to the significant variability in medical terminology, especially between medical literature and day-to-day clinical practice. For example, a specific entity can be lexicalized in many different ways, as is the case with "adriamycin", "doxorubicin", and "hydroxydaunorubicin", all referring to the same drug. Another important consideration is the extensive use of acronyms (such as "chronic obstructive pulmonary disease" and "COPD"), and the fact that the same acronym does not always have a unique designation; "PCR" can refer to "cardiopulmonary resuscitation", "Polymerase Chain Reaction", or "C-Reactive Protein". Finally, as seen in the examples, biological entities can also have multi-word names, further complicating the issue by the need to determine the boundaries of the name and resolve the overlap of candidate names.

Although LLMs demonstrate strong performance in various medical tasks, such as medical QA and summarization, they still fall short of **Pretrained Language Models (PLM)** in performing fundamental but essential tasks, such as clinical NER and identification of clinical record sections (Li et al., 2024). (Hu et al., 2024) show for English, where LLMs are more performant than in some other languages, that GPT-4 achieved an F1 score of 0.861 on certain clinical datasets, whereas BioClinicalBERT (Alsentzer et al., 2019) scored 0.901, suggesting domain-specific models currently maintain an edge in clinical NER. Furthermore, the computational resources needed even for LLM inference exceed what is typically available to hospital professionals.

Unfortunately, obtaining the required ample domain-specific data for the pretraining of PLMs proves to be a challenge, especially within the medical domain. However, it should be noted that significant strides have been made in recent years in improving open-source medical literature data accessibility within the English-speaking community (i.e., National Library of Medicine (2024), National Center for Biotechnology Information (2024)). The same is not true for real clinical data. Due to the sensitive nature of these data, access is tightly controlled and subject to rigorous regulations. This scarcity of clinical data becomes even more pronounced when addressing non-English languages, including Spanish.

Although progress has been made in low-resource languages (Nunes et al., 2024), evidence indicates that domain-specific cross-lingual PLMs effectively mitigate performance disparities by leveraging knowledge transfer from high-resource languages (Hu et al., 2020; Gaschi et al., 2023).

Ideally, in the best possible scenario, healthcare institutions would benefit from PLMs that can be deployed in clinical settings without requiring further continual pretraining on institution-specific data or the costly development of large manually annotated in-domain corpora, such as the case researched in Aracena et al. (2023). This poses several research questions for PLM to be used in real healthcare scenarios:

Research Question 1: Do domain-specific PLMs trained solely on publicly available open-source medical data perform on par with PLMs trained on proprietary data from healthcare institutions? How significant are the performance differences, if any, between lower-level lexical tasks, such as **medical entity recognition (MER)**, and higher-level document tasks, like clinical

notes section identification?

Note that medical literature tends to use highly standardized language, whereas clinical notes written by practitioners in daily practice contain spelling errors, non-standard word usage, a high density of abbreviations, and document structures that can vary significantly between hospitals.

Research Question 2: If performance differences emerge, how significantly are the models affected by changes in the size of the annotated corpus? Do PLMs that continue training on proprietary data from healthcare institutions require less annotated data to achieve performance comparable to those trained on public datasets?

Research Question 3: To what extent can a bilingual medical language model perform effectively on medical tasks that are strictly monolingual? Additionally, does cross-lingual knowledge transfer occur effectively, allowing the model to leverage its bilingual training to improve performance across languages?

To try to solve these questions we have built two English-Spanish bilingual BERT-based Language Models named EriBERTa and EriBERTa private. Pretraining the model in both languages allows to take advantage of the cross-lingual transfer as there is considerably more medical data in English than in Spanish.

Our main **research contributions are**:

1.- **EriBERTa:** The first open-sourced clinical English-Spanish LM.

2.- A set of experiments and a detailed analysis of the impact **private versus public PLMs** have on different clinical tasks.

3.- A set of **monolingual and cross-lingual experiments** measuring the quality of EriBERTa and its transfer learning abilities of EriBERTa (public and private) as well as the adequacy of using comparable but non-equivalent English datasets for intermediate fine-tuning.

2 Related Work

To the best of our knowledge, (Labrak et al., 2023) is the only study that compares the performance of a PLM called DrBERT, trained from scratch on public open-source French medical literature with a PLM called ChuBERT, trained from scratch on a French private healthcare system data, and a mix of both evaluating all of them on a separate subset of that private healthcare system data.

The public open-access data included scientific articles and medical texts in French. The private hospital data included clinical data, mainly composed of **Electronic Health Records (EHRs)** and internal medical documentation. All models are built using RoBERTa architecture and (Labrak et al., 2023) built different DrBERT versions depending on the size of the continual training data. They built a $DrBERT_{large}$ version on a 7.4GB dataset, and a $DrBERT_{small}$ version built on the same dataset size as ChuBERT, namely 4 GB. The study evaluated all models on multiple medical NLP tasks on datasets originated in the private hospital. Tasks such as **Named Entity Recognition (NER)**, part-of-speech (POS) tagging, and text classification. According to their experiments, DrBERT performs as well as or even better than ChuBERT on most tasks despite using only public data. Their results suggest that publicly available biomedical texts can be sufficient for training effective medical NLP models.

Our work differs from (Labrak et al., 2023) in different respects. First, EriBERTa has been trained on bigger datasets. EriBERTa contains a significant amount of bilingual medical as well as clinical data, although this data is approximately four times smaller. Additionally, we perform different task finetuning data size ablations to measure the impact that finetuning dataset size has on the results of both models. Our aim is to check if it is necessary to continually pretrain a model with its own datasets and how much data the private health care system needs to annotate to get similar performance in different tasks.

3 Pretraining EriBERTa

The EriBERTa model is a pretrained transformer-based language model designed to enhance the performance of natural language processing tasks in the medical-clinical domain for the Spanish language. It is based on the RoBERTa architecture, which is a popular pretrained model and a variant of BERT (Devlin et al., 2019). EriBERTa uses the same tokenizer and pretraining methods as RoBERTa, as described in the work of Liu et al. (2019). This ensures that EriBERTa has similar performance characteristics to RoBERTa but with the additional advantage

of being fine-tuned on English and Spanish medical-clinical data. We trained two variations of EriBERTa:

EriBERTa Trained exclusively on publicly available medical and clinical corpora. By leveraging diverse open-access datasets, EriBERTa ensures broad applicability while remaining accessible for research and biomedical-clinical use[1].

EriBERTa-private Pretrained on a combination of public datasets and proprietary EHRs from hospitals, allowing for deeper specialization in real-world clinical language.

In this section, we will delve into the details of the pretraining and evaluation procedures utilized in our research. Specifically, we will begin by describing the corpora used for the pretraining and will highlight the differences between the corpora used for both EriBERTa versions. Finally, we will proceed to explain the specific pretraining procedures for the EriBERTa versions.

3.1 Corpora

To pretrain the EriBERTa model, we used a combination of publicly available and private medical-clinical corpora in Spanish and English. The selection of corpora was based on their relevance to the medical domain and the availability of high-quality data. However, acquiring medical and clinical datasets, particularly in the Spanish language, can prove to be a demanding endeavor.

A primary obstacle in procuring medical and clinical corpora lies in the inherent sensitivity of the data. Medical records are repositories of profoundly personal and confidential information, rendering the collection and dissemination of such data for research purposes an intricate task. These intricacies are further compounded when dealing with the Spanish language. In contrast to English, the availability of medical and clinical corpora in Spanish is notably limited, posing a challenge in amassing sufficient data for the pretraining of language models. Moreover, the quality and quantity of accessible Spanish corpora can vary significantly, potentially impacting the performance of pretrained models.

Despite the challenges of gathering medical and clinical corpora, we have successfully assembled and curated a diverse collection of relevant corpora in both Spanish and English to pretrain the EriBERTa model. Table 1 provides more detailed information on the language and size of each corpus used for pretraining.

Lang	Source	No. Words
	Medical Corpus	
ENG	EMEA	12M
	PubMed Abstracts	968.4M
	Clinical Trials	127.4M
ES	EMEA	13.6M
	PubMed	8.4M
	SNOMED-CT	7.2M
	SPACCC	350k
	UFAL	10.5M
	Wikipedia (Med)	5.2M
	♣ Medical Crawler	918M
	Clinical Corpus (EHRs)	
ENG	MIMIC-III	206M
ES	♠ Private Hospital Documents	222M

Table 1: Pretraining corpora, categorized by document type and language. Corpora marked with ♣ were used exclusively for training the publicly available version, while those marked with ♠ were used solely for the private version.

The corpora we used for pretraining EriBERTa are:

MIMIC-III (Johnson et al., 2016) English database on ICU stays of over 40,000 patients between 2001 and 2012. It contains information such as vital sign measurements, laboratory test results, procedures, medications, caregiver notes, and mortality, among others.

EMEA (Tiedemann, 2012) A parallel corpus in English and Spanish consisting of documents from the European Medicines Agency.

ClinicalTrials A set of English documents on clinical studies carried out worldwide[2].

PubMed Contains abstracts and full texts of biomedical literature from multiple NLM literary sources, including MEDLINE[3],

[1] https://huggingface.co/HiTZ/EriBERTa-base

[2] https://clinicaltrials.gov

[3] https://www.nlm.nih.gov/medline/medline_overview.html

PubMed Central[4], and Bookshelf[5]. We used abstracts for English, and abstracts and full texts for Spanish.

SNOMED-CT (I.H.T.S.D.O, 2022) Standardized and multilingual medical vocabulary consisting of more than 300,000 medical concepts, including categories such as body parts, clinical findings, and pharmaceutical/biological products, among others. For this work, the descriptions in Spanish associated with each term were used.

SPACCC (Intxaurrondo, 2018) Spanish corpus created after collecting 1,000 clinical cases from SciELO[6], and categorizing them based on structure and content into those that were similar to actual clinical texts and those that were not suitable for this task.

UFAL (ÚFAL, 2017) Multilingual medical corpus composed of parallel corpora collected over various projects.

Wikipedia Med A Spanish corpus composed of entries collected from Wikipedia, filtered by scope, and cleaned.

♠ **Private Clinical Documents** A set of Spanish clinical narratives from health centers.

♣ **Medical Crawler (Carrino et al., 2021)** A corpus comprising over 3,000 URLs related to Spanish biomedical and health domains, collected through web crawling.

It is important to note that the main difference between the private and public versions of EriBERTa is the type of clinical documents used for pretraining. **EriBERTa-private** was trained using the Private Clinical Documents corpus but not the Medical Crawler corpus. In contrast, the public version of **EriBERTa** was trained using the Medical Crawler corpus but not the Private Clinical Documents corpus. These differences in pretraining corpora may affect the performance of the models on different tasks, particularly those that involve hospital clinical documents.

3.1.1 Balancing the Corpus

As shown in Table 1, the amount of resources in Spanish and English for the private version of EriBERTa is imbalanced (1,313M words in English and 267M in Spanish). This imbalance could result in English having too much weight when generating the tokenizer and pretraining the model, leading to poor performance in Spanish, as reported in Conneau et al. (2020). Notably, the private EriBERTa version that uses the crawler does not suffer from this imbalance and requires no balancing.

To address this issue, we employed the formula (1) proposed in Conneau and Lample (2019). Here, n_i and p_i represent the number of words and the frequency of occurrence of the language i, respectively, and q_i denotes the probability that a word in language i is sampled according to the multinomial distribution. α is a parameter that controls the language sampling rate, with lower values reducing the sampling probability of the most represented languages and increasing the likelihood of those with scarce resources. Based on the studies described in Conneau et al. (2020) for multilingual models with few resources for some languages, we decided to balance the corpus using a parameter of $\alpha = 0.3$.

$$\{q_i\}_{i=1...N} \;;\; q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{where} \quad p_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \tag{1}$$

In the corpus balancing calculations, we decided to omit the section related to the EHR files because it was already balanced: 206M in English and 222M in Spanish. Without considering these files and applying the formulas in (1), we obtain:

$$p_{es} = 0.04 \qquad p_{eng} = 0.96 \tag{2}$$

$$q_{es} = 0.277 \qquad q_{eng} = 0.723 \tag{3}$$

Therefore, to satisfy the probabilities q_{es} and q_{eng} derived in (2), the Spanish corpus must be scaled by 9.5, expanding from 45.4M to 438M tokens, by duplicating the texts.

3.2 Model Pretraining

In this section, we detail the pretraining process of our EriBERTa models, which are based on the RoBERTa-base architecture. We first describe the tokenizer we used, which is the same as RoBERTa, followed by the vocabulary and pretraining details.

[4] https://www.ncbi.nlm.nih.gov/pmc/about/intro/

[5] https://www.ncbi.nlm.nih.gov/books/

[6] https://scielo.isciii.es/scielo.php

Our tokenizer is based on **Byte-Pair Encoding (BPE)**, introduced by Sennrich et al. (2016), which is a data compression technique used to represent a large set of symbols or words using a smaller vocabulary, converting them to subwords, allowing the tokenizer to handle rare or out-of-vocabulary words. We defined a cased vocabulary with a size of 64,000 that was chosen to accommodate the bilingual nature of the models and ensure sufficient coverage of both languages (Conneau et al., 2020).

For pretraining, we used the **Masked Language Modeling (MLM)** objective for all model variants, which involves randomly masking tokens in the input sequence and training the model to predict the original tokens based on the surrounding context (Devlin et al., 2019). To generate the pretraining data, we concatenated the corpora and then split the concatenated text into segments of the maximum input length allowed by each model. We trained the RoBERTa-based EriBERTa models from scratch for 125k steps, with checkpoints saved every 2.5k steps. The model parameters and training hyperparameters are defined in Table 2.

Training Hyperparameters			
Number of Layers	12	*Batch Size (tokens)*	2,083,840
Hidden Size	768	*Weight Decay*	0.0
FFN inner hidden size	3,072	*Max Steps*	125k
Attention heads	12	*Learning Rate Decay*	Linear with warmup
Dropout	0.1	*Adam* ϵ	1e-08
Attention Dropout	0.1	*Adam* β_1	0.9
Warmup Steps	7.5k	*Adam* β_2	0.99
Peak Learning Rate	2.683e-4	*Gradient Clipping*	1

Table 2: Parameter and hyperparameter details for EriBERTa models.

4 Fine-tuning EriBERTa

In this section, we evaluate the performance of EriBERTa across multiple NER tasks in the biomedical domain, covering both English and Spanish. The evaluation is structured into three main parts, each focusing on a distinct aspect of model performance:

Evaluation on Standard Datasets We evaluate EriBERTa's performance on widely used benchmark datasets for medical NER in both Spanish and English. The English evaluation serves as a validation step to ensure that EriBERTa can effectively perform NER tasks in this language, despite it not being its primary focus. In contrast, the Spanish evaluation is more extensive as it directly aligns with the model's core objective. This assessment provides a deeper analysis of EriBERTa's capabilities in Spanish-language medical text processing.

Transfer Learning Evaluation To assess EriBERTa's cross-lingual generalization capabilities, we perform a zero-shot evaluation. This involves fine-tuning the model in one language and evaluating its performance in the other, allowing us to measure its ability to transfer knowledge between English and Spanish without direct exposure to the target language. The results highlight EriBERTa's effectiveness in cross-lingual medical text processing, demonstrating its potential for multilingual applications in clinical NLP.

Evaluation on Real-World Clinical Datasets Beyond standard benchmarks, we evaluate EriBERTa in real clinical settings using private hospital datasets, focusing on medical entity recognition. Additionally, we investigate the model's performance when trained on limited data. This is a common constraint in hospital environments due to strict data privacy regulations and the scarcity of expert clinician annotators.

To ensure a rigorous assessment, we employed the SeqEval library (Nakayama, 2018) to compute precision, recall, and micro F1 scores for each entity type. Additionally, we conducted an extensive hyperparameter optimization process Biewald (2020), employing the validation split, for each model-dataset combination to explore up to twenty hyperparameter configurations per case. The optimization was performed within a predefined search space (Table 3), using the AdamW optimizer while maintaining default settings from HuggingFace's Transformers library for all other parameters (Wolf et al., 2020). After optimization, we fine-tuned the model for 10 epochs per task, repeating the process five times with different random seeds, and evaluated it on the test split.

Batch size	8, 16, 32
Learning rate (LR)	1e-5, 2e-5, 2.5e-5, 3e-5, 5-e5, 7.5e-5
Weight decay (WD)	0.0, 0.01, 0.2, 0.3, 0.5
Warmup ratio (WR)	0.0, 0.2

Table 3: Hyperparameter search space used for model optimization.

Laguage	Dataset	# Instances			# Labels	Description
		Train	Dev	Test		
EN	NCBI-disease (Doğan et al., 2014)	5,424	923	960	1	This corpus was created for disease recognition and concept normalization tasks. In this work, only the entity recognition task has been carried out.
	BC5CDR-disease (Li et al., 2016)	4,560	4,581	4,797	1	Introduced as part of the BioCreative V challenge, the dataset includes PubMed abstracts annotated with entities and chemical-disease relationships. In this case, the disease recognition part was used.
	BC5CDR-chem (Li et al., 2016)	4,560	4,581	4,797	1	Chemical recognition task from the BC5CDR dataset.
	BC4CHEMD (Krallinger et al., 2015)	7,000	7,000	6,000	1	A biomedical NER dataset introduced in the BioCreative IV challenge. It contains PubMed abstracts annotated with a wide variety of chemical mentions, including drugs, chemical compounds, and formulas.
	JNLPBA (Kim et al., 2004)	37,094	3,857	3,857	1	Created from MEDLINE, this corpus contains 2,000 manually annotated abstracts with 5 entity labels: protein, DNA, RNA, cell line, and cell.
ES	CANTEMIST (Miranda-Escalada et al., 2020)	501	500	300	1	Focuses on neoplasm morphology mentions, including cancer types and tissue abnormalities. Annotated for entity recognition and normalization to oncology terminologies.
	CODIESP (Miranda-Escalada et al., 2020)	500	250	250	2	Clinical case reports annotated with diagnoses and procedures. Entities are linked to ICD-10 codes.
	DisTEMIST (Miranda-Escalada et al., 2022)	600	150	250	1	Includes mentions of diseases ranging from chronic conditions to infections. Annotations are normalized to SNOMED CT.
	MEDDOCAN (Marimon et al., 2019)	500	250	250	21	Comprises clinical texts annotated with personal and institutional identifiers. Entities include names, dates, locations, and contact details.
	MEDDOPROF Subtrack 1 (NER) (Lima-López et al., 2021)	990	248	344	3	Targets mentions of job-related information, including medical professions, work activities, and employment status. Focuses on occupations and labor-related terms.
	MEDDOPROF Subtrack2 (Class) (Lima-López et al., 2021)	990	248	344	4	Classifies text segments that mention individuals according to their role in the clinical context, assigning labels such as patient, healthcare professional, family member, or other.
	MedProcNER (Lima-López et al., 2023)	599	150	250	1	Focuses on mentions of medical procedures and interventions. Includes diagnostic, surgical, and therapeutic processes annotated for normalization.
	MultiCardioNER Track 1 (Lima-López et al., 2024)	1,000	258	250	1	Includes annotations of cardiovascular conditions, covering heart diseases and disorders of the circulatory system.
	MultiCardioNER Track 2 (Lima-López et al., 2024)	1,000	258	250	1	Covers pharmacological substances related to cardiology. Annotations include drugs and therapeutic agents for heart-related conditions.
	PharmacoNER Full Cases (Gonzalez-Agirre et al., 2019)	500	250	250	4	Includes mentions of normalizable and non-normalizable drugs, proteins, and ambiguous terms across full clinical texts.
	PharmacoNER Sentences (Gonzalez-Agirre et al., 2019)	8,129	3,787	3,952	4	Same entity scope as Full Cases but annotated at the sentence level.
	SympTEMIST (Lima-López et al., 2023)	595	149	250	1	An annotated dataset for symptoms, signs, and findings across various medical fields.
	ClinAIS (De la Iglesia et al., 2023a,b)	781	127	130	7	The ClinAIS shared task focuses on identifying seven distinct section types within unstructured clinical records written in Spanish.

Table 4: Summary of the datasets used to evaluate EriBERTa across various biomedical tasks.

4.1 Monolingual Evaluation on Public Datasets

To ensure robust and representative evaluation, we curated a rich collection of publicly available datasets spanning multiple clinical subdomains and entity categories. Our benchmark includes 5 English and 13 Spanish datasets, as summarized in Table 4. These resources feature annotations for diverse medical concepts, ranging from diseases and symptoms to procedures, chemicals, and anatomical structures.

While the English evaluation plays a secondary role, it verifies that EriBERTa can effectively handle English-language medical NER tasks despite not being explicitly optimized for this language, ensuring the quality of the transferred knowledge. In contrast, the Spanish evaluation forms the centerpiece of this study, as the model was intentionally developed to enhance Spanish clinical language processing.

To contextualize EriBERTa's performance, we compare it against strong reference models. For English, we use **BioBERT v1.1** (Lee et al., 2020), a well-established domain-specific model pretrained on biomedical literature, known for its high performance in biomedical NER tasks. Although BioALBERT (Naseem et al., 2021) represents another relevant model in the domain, its architectural differences and divergent model sizes limit its suitability as a direct point of comparison. For Spanish, we employ **bsc-bio-ehr-es** (Carrino et al., 2022), the current state-of-the-art model for Spanish clinical text, trained on a large corpus of EHRs and optimized for Spanish medical NLP tasks.

The full results of this evaluation are presented in Table 5[7]. EriBERTa demonstrates strong performance across the board, particularly in Spanish datasets, where it outperforms the state-of-the-art model in most tasks. In English, despite being trained on significantly less English data and not specifically tailored for this language, EriBERTa performs competitively.

4.2 Transfer Learning in Zero-Shot Scenarios Evaluation

To evaluate the transfer learning capabilities of our models in zero-shot scenarios,

[7]The BioBERT results are those reported in Lee et al. (2020), and for consistency, we used the dataset versions provided by the authors.

Dataset	Reference Model	EriBERTa	EriBERTa-private
NCBI-disease	**89.71**	87.31	87.27
BC5CDR-disease	**87.15**	84.95	84.94
BC5CDR-chem	**93.47**	92.32	92.37
BC4CHEMD	**92.36**	89.85	90.26
JNLPBA	**77.49**	76.68	76.44
Average	**88.04**	86.22	86.26
CANTEMIST	83.59±0.37	**85.09±0.22**	84.85±0.44
CODIESP	65.72±0.33	65.80±0.47	**67.22±0.44**
DisTEMIST	74.44±0.48	**75.76±0.68**	75.29±0.65
MEDDOCAN	96.62±0.28	**96.91±0.34**	96.27±0.24
MEDDOPROF Subtrack 1 (NER)	74.19±1.67	**77.53±0.34**	72.76±0.67
MEDDOPROF Subtrack2 (Class)	**69.72±1.85**	69.39±3.88	65.36±0.47
MedProcNER	75.09±0.36	75.89±0.55	**76.04±0.36**
MultiCardioNER Track 1	72.41±0.29	72.40±0.69	**73.08±0.40**
MultiCardioNER Track 2	90.83±0.68	91.17±0.67	**91.21±0.50**
PharmacoNER Full Cases	88.00±0.43	**90.32±0.46**	89.94±0.45
PharmacoNER Sentences	88.05±0.32	89.09±0.68	**89.59±0.11**
SympTEMIST	67.51±0.52	69.59±0.63	**69.77±0.56**
ClinAIS	76.06±1.18	**76.81±0.96**	75.66±0.18
Average	78.64	**79.67**	79.01

Table 5: Performance of EriBERTa on various monolingual biomedical datasets and comparison with the reference models for each language.

we leveraged the DIANN dataset (Fabregat et al., 2018), which provides a bilingual setting where each document is available in both Spanish and English. This dataset focuses on identifying disability-related mentions, such as "low vision" and "deafness". Given the broader availability of annotated datasets in English compared to Spanish, we aimed to assess whether models trained on English data could effectively perform the same task in Spanish without requiring additional Spanish annotations by effectively transferring the learned knowledge.

For this experiment, we fine-tuned EriBERTa on one language and evaluated its performance on the other, using the best hyperparameter configuration identified through optimization and training for 15 epochs instead of 10. This approach allowed us to analyze the extent to which cross-lingual transfer learning can mitigate resource disparities and enhance model deployment in Spanish-language clinical NLP tasks.

The results of this zero-shot evaluation, presented in Table 6, confirm EriBERTa's ability to generalize across languages. Both model variants successfully transferred knowledge and achieved adequate performance in the target language, with particularly promising results in the English-to-Spanish transfer, the primary intended use case. Notably, the public version of EriBERTa demonstrated superior cross-lingual transfer learning capabilities compared to the private version.

Model	Training Language	Target Language: EN	Target Language: ES
EriBERTa	*EN*	80.89±1.66	71.97±2.12
	ES	71.33±4.83	81.78±1.32
EriBERTa Private	*EN*	80.83±0.73	67.12±4.53
	ES	63.06±2.55	79.09±1.73

Table 6: Performance comparison of EriBERTa in a zero-shot transfer learning scenario, evaluated on English and Spanish.

We also compared its performance with the best model presented by Goenaga et al. (2023). In their study, they assessed the performance of the XLM-RoBERTa (Conneau et al., 2020) model and multiple FLAIR (Akbik et al., 2019) approaches in the Spanish zero-shot case. The comparative analysis, presented in Table 7, demonstrates that EriBERTa surpasses their results.

System	*Precision*	*Recall*	*Micro F1-Score*
FLAIR $_{ME}$	50.64	34.50	41.04
EriBERTa	**68.46**	**76.24**	**71.97**
EriBERTa-private	67.51	66.81	67.12

Table 7: Comparison of the EriBERTa models and the state-of-the-art system trained only on English and evaluated in Spanish.

4.3 Evaluation in Real-World Clinical Datasets

Beyond standard benchmarks, we assess EriBERTa's performance in real-world clinical settings using private hospital datasets. This evaluation focuses on two key tasks: **medical entity recognition (MER)** and **clinical section classification**. Additionally, we analyze the models' robustness in data-scarce scenarios, simulating common constraints in hospital environments due to stringent data privacy regulations and the limited availability of expert clinician annotators. By systematically reducing the training data, we aim to compare the adaptability of both EriBERTa versions under varying levels of data scarcity.

4.3.1 Datasets

We utilized three datasets from the same healthcare center that provided the EHRs used to pretrain EriBERTa. It is important to note that the EHRs used for pretraining were not included in the validation or test splits of these datasets, ensuring an unbiased evaluation of the model's performance.

For MER, we use a refined and expanded version of the dataset introduced by Casillas et al. (2019), which annotates seven distinct medical entity types[8] within clinical documents. The dataset consists of 434 EHRs for training, and 92 each for validation and test.

For **clinical section classification**, we employ two variations of the dataset presented in Goenaga et al. (2021), which consists of semi-structured clinical notes. The goal is to correctly classify different segments of each document into predefined sections, such as "Present Illness", "Exploration", and "Treatment". One dataset variant contains non-standardized section headers, while the other lacks section headers entirely, providing a challenging test of the models' ability to infer structure from unstructured text. The dataset consists of 300 clinical notes, evenly split into training, validation, and test sets with 100 notes each.

4.3.2 Experimental Setup

To systematically assess the impact of data scarcity, we conduct experiments by training the models on progressively smaller subsets of the dataset, using 5%, 10%, 20%, 30%, 50%, 75%, and 100% of the available data. For each subset size, we perform the hyperparameter optimization process and then five independent runs with the best hyperparameter values. To ensure robustness and mitigate selection bias, a different random subset is selected for each independent run. This approach enables a comprehensive comparison of the generalization capabilities of both EriBERTa versions under different data availability conditions and analyzes the impact of pretraining a model with clinical documents from the same hospital.

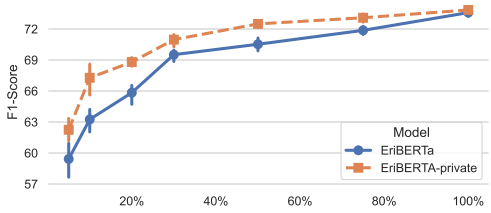
4.3.3 Results

The results, summarized in Table 8, underscore the clear advantage of leveraging in-domain clinical data during pretraining. The EriBERTa-private model, which was pretrained on EHRs originating from the same hospital as the downstream evaluation datasets, consistently surpasses the performance of its publicly pretrained counterpart. This performance gap is especially pronounced in the clinical section classification task, where EriBERTa-private benefits from its prior exposure to documents with similar lexical, structural, and stylistic characteristics.

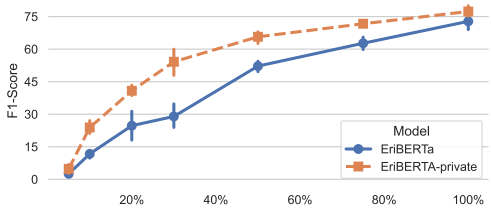
Task	EriBERTa	EriBERTa-private
Medical Entity Recognition (MER)	73.59±0.42	**73.84±0.34**
Section Detection	72.80±5.00	**77.35±3.15**
Section Detection *[w/o Headers]*	52.49±1.96	**59.33±3.02**

Table 8: Evaluation results on private hospital datasets for two crucial tasks.

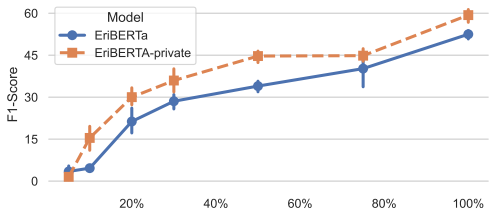
Crucially, this advantage becomes even more evident under low-resource conditions. As shown in Figure 1, EriBERTa-private demonstrates significantly greater robustness when fine-tuned with limited annotated data.



(a) Medical Entity Recognition.



(b) Section Detection.



(c) Section Detection without Headers.

Figure 1: Performance comparison of EriBERTa and EriBERTa-private under varying levels of training data availability.

[8]Drugs, Diseases, Body Structures, Allergies, Procedures, and Qualifiers.

5 Discussion

This study investigates the effectiveness of domain-specific PLMs in clinical natural language processing, with a particular focus on the impact of in-domain data, data scarcity, and multilingual capabilities. We structured our analysis around three core research questions, exploring performance differences between public and proprietary pretraining data, the impact of annotated dataset size, and the potential for cross-lingual transfer in bilingual models. In the following sections, we discuss our findings concerning each question, highlighting their implications for both clinical NLP development and real-world deployment.

RQ1 - Public vs. Proprietary Pretraining Our results show that PLMs trained on public medical corpora can achieve strong performance across a variety of clinical NLP tasks. As shown in Subsection 4.1, EriBERTa demonstrated competitive results in MER and section classification, even outperforming established reference models in Spanish and matching strong baselines in English.

However, the benefits of training on proprietary clinical notes from the same institution become evident when models are evaluated on tasks involving real-world hospital narratives, as verified in Subsection 4.3. In this setting, EriBERTa-private consistently outperformed its public counterpart. This performance gap was especially pronounced in higher-level tasks such as section classification, which are highly sensitive to document structure and local clinical conventions.

These findings suggest that while publicly sourced medical corpora enable robust general-purpose modeling, incorporating institution-specific data can yield significant gains, especially in tasks that require deep contextual understanding of clinical documentation practices unique to a particular healthcare setting.

RQ2 - Data Efficiency Under Scarcity As shown in Subsection 4.3, EriBERTa-private demonstrates a clear advantage in low-data regimes, achieving strong performance with significantly less annotated data. Moreover, its performance continues to improve steadily as more training data becomes available, highlighting superior scalability. This effect is especially pronounced in the clinical section classification task, which relies heavily on understanding the structure and conventions of documentation specific to a given healthcare institution. In contrast, medical entity recognition tasks, being more lexical and standardized, show smaller gains under the same conditions.

RQ3 - Bilingual and Cross-Lingual Capabilities As shown in Subsection 4.1 and discussed in RQ1, both EriBERTa models, despite being bilingual, achieve strong performance on strictly monolingual tasks, remaining competitive in English and outperforming the state of the art in Spanish. In Subsection 4.2, we further demonstrate their capacity for effective cross-lingual knowledge transfer. This ability is especially valuable for leveraging annotated data from high-resource languages to support tasks in lower-resource ones, such as Spanish, where annotated clinical datasets are scarce.

6 Conclusions

In this work, we presented EriBERTa, the first open-source bilingual clinical language model for English and Spanish, designed to bridge the gap in medical NLP for under-resourced languages and settings. Through a comprehensive series of experiments, we evaluated the impact of pretraining on public versus proprietary clinical data, the model's performance under varying data availability conditions, and its ability to transfer knowledge across languages.

Our results confirm that while domain-specific PLMs trained solely on public data can perform competitively on general clinical tasks, models pretrained on EHRs from the same institution as the target data exhibit significant performance gains, particularly in tasks that heavily rely on institutional document structures or conventions. This advantage becomes even more evident in low-resource settings. Furthermore, we show EriBERTa's robustness in monolingual tasks and knowledge transfer across languages in zero-shot scenarios. This cross-lingual capability is especially relevant in the clinical domain, where annotated datasets in many languages remain scarce.

Additionally, by releasing EriBERTa publicly, we provide the community with a resource to advance multilingual medical NLP and foster further research in cross-lingual, low-resource, and institution-specific clinical settings.

Acknowledgements

This work has been partially supported by the HiTZ Center and the Basque Government (IXA excellence research group funding IT-1570-22 and IKER-GAITU project), as well as by the Spanish Ministry of Universities, Science and Innovation MCIN/AEI/10.13039/501100011033 by means of the projects: Proyectos de Generación de Conocimiento 2022 (EDHER-MED/EDHIA PID2022-136522OB-C22), DeepKnowledge (PID2021-127777OB-C21), DeepMinor (CNS2023-144375), and EU NextGeneration EU/PRTR (DeepR3 TED2021-130295B-C31). And also by an FPU grant (Formación de Profesorado Universitario) from the Spanish Ministry of Science, Innovation and Universities (MCIU) to the first author (FPU23/03347).

References

2015. Connecting health and care for the nation: A shared nationwide interoperability roadmap. Office of the National Coordinator for Health Information Technology (ONC). Washington, DC: U.S. Department of Health and Human Services (HHS).

2019a. Health Level Seven (HL7). CDA. `http://www.hl7.org`. Last Online; accessed 31-05-2021.

2019b. Health Level Seven (HL7). FHIR. `http://www.hl7.org`. Last Online; accessed 31-05-2021.

2019. Recommendation on a European Electronic Health Record exchange format. European Commision.

2020. Health at a Glance: Europe 2018 STATE OF HEALTH IN THE EU CYCLE. `https://ec.europa.eu/health/sites/default/files/state/docs/2018_healthatglance_rep_en.pdf`. Last Online; accessed 31-05-2021.

2020. openehr. `https://www.openehr.org`. Last Online; accessed 31-05-2021.

2020. State of Interoperability among U.S. Non-federal Acute Care Hospitals in 2018 . `https://www.healthit.gov/sites/default/files/page/2020-03/State-of-Interoperability-among-US-Non-federal-Acute-Care-Hospitals-in-2018.pdf`. Last Online; accessed 31-05-2021.

. 2021. Health IT Data Summaries. `https://dashboard.healthit.gov/apps/health-information-technology-data-summaries.php`. Last Online; accessed 31-05-2021.

. 2022. International Statistical Classification of Diseases and Related Health Problems (ICD). `https://www.who.int/standards/classifications/classification-of-diseases`. Last Online; accessed 31-05-2022.

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alsentzer, E., J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Aracena, C., N. Rodríguez, V. Rocco, and J. Dunstan. 2023. Pre-trained language models in Spanish for health insurance coverage. In T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 433–438, Toronto, Canada, July. Association for Computational Linguistics.

Biewald, L.. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Carrino, C. P., J. Armengol-Estapé, O. de Gibert Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, and M. Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical*

Language Processing, pages 193–199, Dublin, Ireland, May. Association for Computational Linguistics.

Casillas, A., N. Ezeiza, I. Goenaga, A. Pérez, and X. Soto. 2019. Measuring the effect of different types of unsupervised word representations on medical named entity recognition. *International Journal of Medical Informatics 129*, 100–106.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Conneau, A. and G. Lample. 2019. Cross-lingual language model pretraining. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

De la Iglesia, I., M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, and A. Atutxa. 2023a. An Open Source Corpus and Automatic Tool for Section Identification in Spanish Health Records. *Journal of Biomedical Informatics 145*, 104461.

De la Iglesia, I., M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, and A. Atutxa. 2023b. Overview of ClinAIS at IberLEF 2023: Automatic Identification of Sections in Clinical Documents in Spanish. *Procesamiento del Lenguaje Natural 71*, 289–299.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Doğan, R. I., R. Leaman, and Z. Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics 47*, 1–10.

Fabregat, H., J. Martínez-Romo, and L. Araujo. 2018. Overview of the DIANN task: Disability annotation task. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) colocated with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, Volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

Gaschi, F., X. Fontaine, P. Rastin, and Y. Toussaint. 2023. Multilingual clinical ner: Translation or cross-lingual transfer? In *5th Clinical Natural Language Processing Workshop*, pages 289–311. Association for Computational Linguistics.

Goenaga, I., E. Andres, K. Gojenola, and A. Atutxa. 2023. Advances in Monolingual and Crosslingual Automatic Disability Annotation in Spanish. *BMC Bioinformatics*.

Goenaga, I., X. Lahuerta, A. Atutxa, and K. Gojenola. 2021. A section identification tool: Towards hl7 cda/ccr standardization in spanish discharge summaries. *Journal of Biomedical Informatics 121*, 103875.

Gonzalez-Agirre, A., M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, and M. Krallinger. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China, November. Association for Computational Linguistics.

Hu, J., S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Hu, Y., Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering.

I.H.T.S.D.O. 2022. *SNOMED CT - Starter Guide*. Online: International Health Terminology Standards Development Organisation.

Institute of Formal and Applied Linguistics (ÚFAL). 2017. UFAL: Medical Corpus. `https://ufal.mff.cuni.cz/ufal_medical_corpus`.

Intxaurrondo, A.. 2018. Spaccc. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data 3*, 160035.

Kim, J.-D., T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, pages 70–75, USA. Association for Computational Linguistics.

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. lu, R. Leaman, Y. Lu, D. Ji, D. Lowe, R. Sayle, R. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, W. Qi, and A. Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics 7*, S2.

Labrak, Y., A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, and P.-A. Gourraud. 2023. DrBERT: A Robust Pretrained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada, July. Association for Computational Linguistics.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform. 36*(4), 1234–1240.

Li, J., Y. Deng, Q. Sun, J. Zhu, Y. Tian, J. Li, and T. Zhu. 2024. Benchmarking large language models in evidence-based medicine. *IEEE Journal of Biomedical and Health Informatics*.

Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford) 2016*.

Lima-López, S., E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret, and M. Krallinger. 2023. 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*.

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

Lima-López, S., E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, and M. Krallinger. 2021. NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural 67*, 243–256.

Lima-López, S., E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. 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 G. Faggioli,

N. Ferro, P. Galuščáková, and A. García Seco de Herrera, editors, *CLEF Working Notes*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*.

López-Úbeda, P.. 2022. Biomedical entities recognition in Spanish combining word embeddings. *Proces. del Leng. Natural 68*, 149–152.

Marimon, M., A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@ SEPLN*, 303–323.

Miranda-Escalada, A., L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

Miranda-Escalada, A., A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger. 2020. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. In L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, Volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nakayama, H.. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Naseem, U., M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim. 2021. BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–7. IEEE.

National Center for Biotechnology Information. 2024. NCBI Databases and Tools. U.S. National Library of Medicine, Accessed April 3, 2025.

National Library of Medicine. 2024. PubMed Database. U.S. National Library of Medicine, Accessed April 3, 2025.

Nunes, M., J. Boné, J. C. Ferreira, P. Chaves, and L. B. Elvas. 2024. MediAlbertina: An european portuguese medical language model. *Computers in Biology and Medicine 182*, 109233.

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Tiedemann, J.. 2012. Parallel data, tools and interfaces in opus. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 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, October. Association for Computational Linguistics.