Artificial Intelligence In Medicine 143 (2023) 102625

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Transformers for extracting breast cancer information from Spanish clinical narratives

Oswaldo Solarte-Pabón [a,b,*], Orlando Montenegro [b], Alvaro García-Barragán [a], Maria Torrente [c], Mariano Provencio [c], Ernestina Menasalvas [a], Víctor Robles [a]

[a] Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain
[b] Escuela de Ingeniería de Sistemas, Universidad del Valle, Cali, Colombia
[c] Hospital Universitario Puerta de Hierro de Madrid, Madrid, Spain

ARTICLE INFO

Keywords:
Natural Language Processing (NLP)
Named Entity Recognition (NER)
Deep learning
Breast cancer
Clinical narratives

ABSTRACT

The wide adoption of electronic health records (EHRs) offers immense potential as a source of support for clinical research. However, previous studies focused on extracting only a limited set of medical concepts to support information extraction in the cancer domain for the Spanish language. Building on the success of deep learning for processing natural language texts, this paper proposes a transformer-based approach to extract named entities from breast cancer clinical notes written in Spanish and compares several language models. To facilitate this approach, a schema for annotating clinical notes with breast cancer concepts is presented, and a corpus for breast cancer is developed. Results indicate that both BERT-based and RoBERTa-based language models demonstrate competitive performance in clinical Named Entity Recognition (NER). Specifically, BETO and multilingual BERT achieve F-scores of 93.71% and 94.63%, respectively. Additionally, RoBERTa Biomedical attains an F-score of 95.01%, while RoBERTa BNE achieves an F-score of 94.54%. The findings suggest that transformers can feasibly extract information in the clinical domain in the Spanish language, with the use of models trained on biomedical texts contributing to enhanced results. The proposed approach takes advantage of transfer learning techniques by fine-tuning language models to automatically represent text features and avoiding the time-consuming feature engineering process.

1. Introduction

Cancer remains one of the main public health problems, ranked as the leading cause of death globally [1]. According to the World Health Organization,[1] cancer caused nearly 10 million deaths worldwide in 2020. In 2022, 1,918,030 new cancer cases and 609,360 cancer deaths are projected to occur in the United States. In particular, breast cancer is currently the most common cancer globally, accounting for 12.5% of all new annual cancer cases worldwide.[2] In 2023, an estimated 297,790 women in the United States will be diagnosed with invasive breast cancer.[3]

The process of diagnosing and treating cancer patients generates a huge amount of information that describes symptoms, the cancer diagnosis, family history, treatments, and the evolution of the patient at the time [2,3]. Physicians register this information in Electronic Health Records (EHR) using clinical notes written in narrative form [4]. Extracting and mining this information is crucial to support oncology research, design treatment plans, and improve patient outcomes [5]. This has been the main goal of two European projects: IASIS[4] and CLARIFY[5] funded by H2020. One of the goal of this projects was to understand risk factors for cancer and describe patterns for progression and relapse. In order to be able to generate these models structured information for each patient is required.

Extracting information from clinical narratives is a challenge due to the complexity of natural language [6]. Moreover, clinical texts are written by highly skilled physicians and nurses using domain-specific terms, under time pressure, with rich and complex jargon, which makes these texts differ from those of other domains [7]. In recent years the use of Natural Language Processing (NLP) in the biomedical domain

* Corresponding author at: Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain.
E-mail address: oswaldo.solartep@alumnos.upm.es (O. Solarte-Pabón).

[1] https://www.who.int/news-room/fact-sheets/detail/cancer.
[2] https://www.breastcancer.org/facts-statistics.
[3] https://www.breastcancer.org/facts-statistics.
[4] https://project-iasis.eu/.
[5] https://www.clarify2020.eu/.

https://doi.org/10.1016/j.artmed.2023.102625
Received 20 December 2022; Received in revised form 11 May 2023; Accepted 8 July 2023
Available online 13 July 2023

has increased the possibility of automatically extracting information from clinical narratives [8–10]. The application of NLP and Artificial Intelligence (AI) techniques for processing medical records plays an increasingly significant role in advancing clinical decision support [11]. The use of EHR to perform studies in the cancer field has also increased in the last few years [12].

The first challenge to be addressed when extracting information from clinical texts is the identification of medical-named entities. Extracting named entities is one of the most important tasks in the medical domain since performing clinical studies commonly requires detailed patient information recorded in clinical notes [6,13]. Clinical Named Entity Recognition (Clinical NER) is the task that aims to identify medical concepts from clinical text [14,15]. In the oncology domain, the first rule-based approaches focused on the extraction of the cancer tumor stage [16–18]. The main limitation of rule-based approaches is they can suffer from a lack of flexibility and universality [19]. Other studies have used machine learning-based approaches to extract other cancer-related concepts such as cancer diagnosis [20], cancer symptoms [21], or treatments [22]. A disadvantage of these approaches is they require a considerable set of text features which frequently are human-dependent and are time-consuming [23].

Recently, deep learning-based approaches have shown important advances and improvements in extracting information in the biomedical domain [24–27]. However, most of these proposals have focused on the English language [24,28]. In fact, information extraction in the medical domain represents its own challenges in languages other than English [29,30]. In the case of the Spanish language, previously performed studies only support extracting a limited set of medical concepts and do not offer a comprehensive information model in the cancer domain [31,32]. These facts indicate that there is a lack of computational models to support Clinical NER in the breast cancer domain for the case of the Spanish language.

Motivated by the success of deep learning for processing natural language texts, in this paper, we propose a transformer-based approach to extract named entities from breast cancer clinical notes written in Spanish. The main contributions of the proposed approach are the following:

- A breast cancer corpus that oncology experts have manually annotated. This in-house corpus is based on a comprehensive annotation schema that contains 23 labels to represent medical concepts in the breast cancer domain. To the best of our knowledge, this is the first corpus aiming to support breast cancer entity extraction in the Spanish language.
- Transformer-based models to perform clinical NER as a sequence-labeling task. These models have been trained on the breast cancer corpus using BERT [33] and RoBERTa [34]. We chose BERT and RoBERTa because these transformers models have been widely used to perform sequence-labeling tasks in the medical domain [30,35,36]. @ Moreover, they can learn long-range dependencies between words using a self-attention mechanism [37]. This mechanism can be useful for learning dependencies between different labels in the breast cancer corpus. Finally, both BERT and RoBERTa contain multilingual pre-trained versions which offer the possibility of training models for the Spanish language.
- The proposed approach takes advantage of transfer learning techniques by fine-tuning transformer models to automatically represent text features, thus avoiding the time-consuming feature engineering process. Therefore, we conduct and compare a fine-tuning process using four distinct monolingual and multilingual models pre-trained on BERT and RoBERTa, including BETO [38], Multilingual BERT [33], RoBERTa Biomedical [39], and RoBERTa BNE [40]. The main characteristics of these models are shown in Table 3. Results obtained have shown competitive performance indicating that the proposed approach is feasible to perform clinical NER from clinical texts in the breast cancer domain.
- Once transformer-based models are obtained and validated, they can be used to extract information from clinical records structuring the information of each patient. This information is used to build models to predict cancer recurrences or progression, among others. For instance, the Projects IASIS[6] and CLARIFY[7] financed by "H2020 programme" used clinical NER models to find risk factors for breast and lung cancer. Electronic Health records from a hospital in Madrid region were the source of data.

The remainder of this document has been organized as follows: Section 2 shows a review of relevant studies that aim to extract medical concepts in the cancer domain. Section 3 describes the proposed approach in this paper. Section 4 presents the results obtained in experiments, and Section 5 provides a discussion of these results. Finally, Section 6 presents the main conclusions and outlook for future work.

2. Background

In recent years the use of Natural Language Processing (NLP) in the biomedical domain has increased the possibility of automatically extracting information from clinical narratives [8,9,14]. In particular, **Clinical Named Entity Recognition** (Clinical NER) is the task that aims to identify medical concepts from clinical text [14,15]. @Clinical NER can also be referred to as Clinical concept extraction [41]. Extracting named entities is one of the most important tasks in the medical domain since performing clinical studies commonly requires detailed patient information recorded in clinical narratives [6,13].

The use of NLP-based approaches to perform clinical NER in the cancer domain has grown recently since extracted information is crucial to perform evidence-based medicine, health quality improvement, and create patient-centered treatments [42]. In this section, we will describe relevant studies conducted to extract clinical named entities in the cancer field. Clinical NER has been commonly addressed using Rules, Machine learning, and Deep learning-based approaches, as follows:

2.1. Rules

The first studies to extract medical concepts in the cancer field used rule-based approaches mainly to obtain the cancer tumor stage [16, 43,44]. The cancer stage indicates the grade and size of the tumor when the patient was diagnosed [17,18,45]. Cancer stage identification has a significant value in predicting incidence and mortality in cancer patients [46]. The proposal described in [47] uses heuristic rule-based algorithms for extracting three variables related to liver cancer staging: tumor size, staging level, and the percentage of the liver invaded by the tumor.

The study carried out in [32] uses a set of regular expressions and the UMLS[8] dictionary to extract several cancer concepts such as tumor stage, mutation status, and the patient performance status from lung cancer clinical notes. @The main weakness of this proposal is that they rely on hand-crafted rules and are limited by the dictionary of medical terms. Rule-based approaches fail when extracting medical concepts with high data variability since they lack flexibility [19]. Moreover, identifying only the cancer stage is not sufficient to have a comprehensive understanding of cancer behavior from clinical narratives.

[6] https://project-iasis.eu/.
[7] https://www.clarify2020.eu/.
[8] https://www.nlm.nih.gov/research/umls/index.html.

2.2. Machine learning

Machine learning-based approaches extract information using annotated data and a set of defined features. In these approaches, the extraction of cancer concepts is defined as a classification problem where each token in a sentence is tagged with a pre-defined label. Machine learning-based approaches work better for tasks for which the set of extraction concepts is large, and they have high data variability [21]. Various algorithms, including Support Vector Machines (SVM), [48] and Conditional Random Fields (CRF) [49], have been widely used to extract cancer information from clinical narratives.

In [21], the authors propose a CRF-based approach to extract breast cancer symptoms using three labels: symptom (positive label), absence of a symptom (negative label), or no symptom at all (neutral label). Features at the work level, such as lower-case words, the last three characters of the word, and the last two characters of the word are used to train the model. @Another CRF-based approach to extract symptoms in cervical cancer is described in [50]. This approach uses the Stockholm EPR corpus, a database containing a Swedish clinical text dataset.

In [51,52], the authors use the SVM algorithm to extract colorectal cancer diagnosis and the cancer stage, respectively. In [53] are described two machine learning-based models to extract information from pathology reports also using the Support Vector Machine (SVM) algorithm. The first model is used to classify notes into internal (primary review) and external (consultation) reports. The second model was used to extract dates and tumor location in patients diagnosed with gastroesophageal cancer. The main disadvantage of the above approaches is they require a considerable set of hand-crafted features that depend on humans and are time-consuming [23].

2.3. Deep learning

In recent years, deep learning-based approaches have shown significant progress, and improvements in extracting information related to cancer [10,54,55]. The main advantage of deep learning approaches is the ability to automatically learn high-level features from texts, reducing the time in the hand-crafted feature engineering process. Moreover, using deep learning methods has opened the opportunity to extract comprehensive sets of medical concepts related to cancer [55,56].

One of the first studies to extract cancer information using a deep learning-based approach is described in [24]. This study aims to extract lung cancer stages, histology, tumor grades, and therapies (chemotherapy, radiotherapy, surgery) using convolutional neural networks (CNN). The authors highlight the feasibility of extracting cancer-related information from clinical narratives using deep learning methods. @In [28], the authors proposed a Bidirectional Long Short Memory (BiLSTM) neural network for extracting radiotherapy treatments from clinical narratives written in English. This study extracts detailed information related to radiotherapy treatment, such as dosage, frequency, fraction frequency, and treatment site.

The use of deep-learning methods has also encouraged the extraction of more detailed information related to cancer. For instance, in [56], the authors described a deep-learning approach to extract breast cancer concepts using BERT. The goal of this proposal is to extract a comprehensive set of breast cancer concepts from clinical notes written in Chinese. The authors demonstrate that the BERT-based model performs better than traditional machine learning algorithms at extracting named entities in the cancer field. @In [57], the authors describe a BiLSTM-based model for clinical concept extraction from oncological clinical notes written in German. This model supports extracting several concepts such as diagnosis, treatments, and medications. Although deep learning-based approaches have improved the ability to extract medical concepts in the cancer medical field, most of these proposals have focused on the English language [24,28] and most recently, on Chinese [55,56].

In the Spanish language case, in [31], the authors propose Cantemist, an annotated corpus to support tumor morphology extraction. @Several studies [58–60] have used this corpus to perform morphology extraction. However, the main limitation of these proposals is they only support identifying one entity type (tumor morphology). Cancer is a complex and specialized medical field requiring a comprehensive set of medical concepts for understanding its evolution from clinical narratives [56]. Extracting entities such as the cancer diagnosis, treatments, and family history is crucial to understand the cancer evolution in the patient. Moreover, named entities such as cancer biomarkers and comorbidities are also important to understand the cancer evolution in the patient. Cancer biomarkers refer to biological molecules produced by the body or tumor in a patient [61]. Knowing cancer biomarkers is useful because physicians can use them for defining alternative treatment plans [62]. On the other hand, comorbidities refer to another illness (e.g., diabetes mellitus, dyslipidemia, arterial hypertension) that co-exists together with a cancer diagnosis. Extracting and analyzing these comorbidities is an important step because they affect the diagnosis and evaluation of treatment effectiveness [63].

Table 1 shows a summary of the most relevant approaches aimed at performing clinical NER in the cancer medical field. From this table, it is important to highlight the following facts:

- Most of the proposals have focused on the English language. According to Table 1, close to 70% of the approaches have concentrated on English.
- Most of the studies have focused on a reduced set of medical concepts (e.g., tumor stage, tumor morphology). Although the proposal described in [56] aimed to extract a comprehensive set of medical concepts, that proposal has focused on the Chinese language.
- Extracting named entities from oncology clinical texts written in Spanish has not been explored deeply yet. There is a lack of corpora to support information extraction in the breast cancer domain in this language.

3. Materials and methods

Transformers for extracting breast cancer information from clinical narratives will be used. Transformers can learn long-range dependencies between words using a self-attention mechanism [37]. In this approach, clinical NER is performed as a sequence labeling task, where each token in a sentence is classified with a specific label. Fig. 1 shows the proposed approach, which consists of three steps: (i) Corpus generation, (ii) Model training, and (iii) Model validation.

3.1. Corpus generation

We randomly chose 500 clinical narratives belonging to breast cancer patients from a public hospital in Madrid, Spain. These clinical notes were anonymized, split into sentences, and tokenized. The BRAT[9] tool was used for the manual annotation process. In what follows, we detail the process of annotating these notes to generate an annotated corpus. The process includes the following steps: definition of the annotation schema, annotating clinical notes, and measuring the reliability of the annotations

[9] https://brat.nlplab.org/.

Table 1
Related works for extracting cancer named entities.

Proposal	Approach	Method	Target	Lang
Nguyen et al. 2010 [64]	Rules	Regex, Dictionaries	Cancer stage	English
Yim et al. 2016 [47]	Rules	Heuristics	Cancer stage	English
Warner et al. 2016 [43]	Rules	Regex	Cancer stage	English
Soysal et al. 2019 [44]	Rules, Machine learning	Regex, Dictionaries	Cancer stage biomarkers	English
Najabadipour et al. 2018 [32]	Rules	Regex, Dictionaries	Cancer stage	Spanish
Forsyth et al. 2018 [65]	Machine learning	CRF	Symptoms	English
Weegar et.al,,2015 [50]	Machine learning	CRF	Symptoms	Swedish
Lenain et al. 2019 [52]	Machine learning	SVM	Diagnosis	English
Huang et al. 2015 [66]	Machine learning	SVM	Cancer stage	English
Wang et al. 2019 [24]	Deep learning	CNN	Stage, histology, Tumor grades, Therapies	English
Bitterman et al, 2020 [67]	Deep learning	BiLSTM	Radiotherapies	English
Zhang et al, 2019 [56]	Deep learning	BERT	Comprehensive model	Chinese
Kittner et al. 2021 [57]	Deep learning	BiLSTM	Diagnosis, treatments, medications	German
Lopez-Ubeda et al. 2020 [60]	Deep learning	BiLSTM	Tumor morphology	Spanish

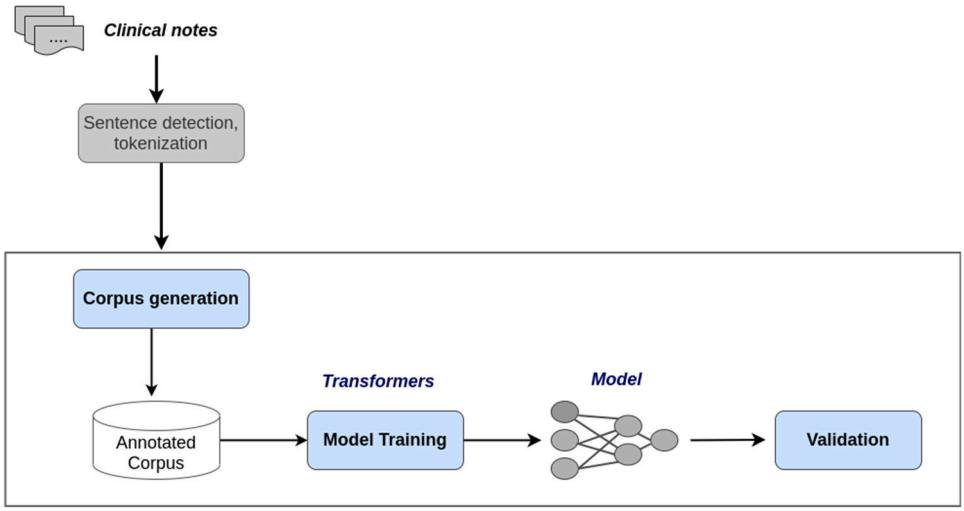

Fig. 1. Proposed approach for extracting breast cancer information.

3.1.1. Definition of the annotation schema

The main goal of this step is to identify the most relevant entities that describe the breast cancer domain. The experts in the breast cancer domain define this set of entity types. Fig. 2 shows a set of sentence examples and their annotations in the cancer corpus. From this Figure, we can highlight that the main challenges for extracting named entities from clinical notes of cancer patients are related to the following facts:

- **Rich set of semantic types:** oncological clinical notes contain a significant number of medical concepts used by physicians to describe the evolution of cancer on the patient including cancer diagnosis, cancer treatments, patient comorbidities, cancer biomarkers, drug names, dates and others.
- **Family history:** it is an important fact because some clinical concepts are related to the patient, but others are related to the family. Thus, a clinical IE system requires to differentiate between what concepts are related to the patient and what are not.
- **Different length entities**: clinical notes contain entities composed of many tokens (lengthy entities) and also entities composed of one token. Lengthy entities are common in medical concepts such as the cancer diagnosis where the physician writes the most specific tumor description. For instance, in Fig. 2 the cancer concept (*"carcinoma ductal infiltrante de mama izquierda"*) can be considered a lengthy entity. Thus, in these cases, a challenge is to extract exactly all the tokens contained in the medical concept. Other concepts such as drug names frequently contains one token.
- **Polysemy**: it is also a common linguistic phenomenon that appears in clinical notes of cancer patients. This occurs when the same word has different meanings depending on its context. From Fig. 2, we can see that the token *DL* in the first line refers to a comorbidity. On the other hand, the token *dL* in the line number 14 (*"Analiticas cortisol, ug/dL"*), is part of a metric for a medical test. These tokens have different meanings since their context are also different.
- **Frequency of medical concepts**: in oncological clinical notes, there are medical concepts that are mentioned by physicians frequently. This is the case of cancer concepts, drug names, dates, and time expressions. In contrast, other important medical concepts such as family members, or toxic habits may appear less frequently.
- **Variability:** it means the number of different instances of a semantic type. The variability in that medical concepts appear is different in clinical cancer notes. There are medical concepts such as family members, or toxic habits with low variability. But also, there are other concepts such as cancer diagnosis with high variability. Thus, correctly extracting named entities with different variability is also a challenge.

To address the above facts, we propose an annotation schema composed of twenty three entity types that are described in what follows (Fig. 2 shows a set of sentence examples and their annotations):

- **Cancer entity** is used to identify tumors description mentioned by clinicians (e.g., carcinoma, adenocarcinoma, cancer). This description also can include the anatomical location of the tumor (e.g., right breast carcinoma). The cancer entity is also used for annotating metastasis concepts (e.g., liver metastasis, bone metastasis).

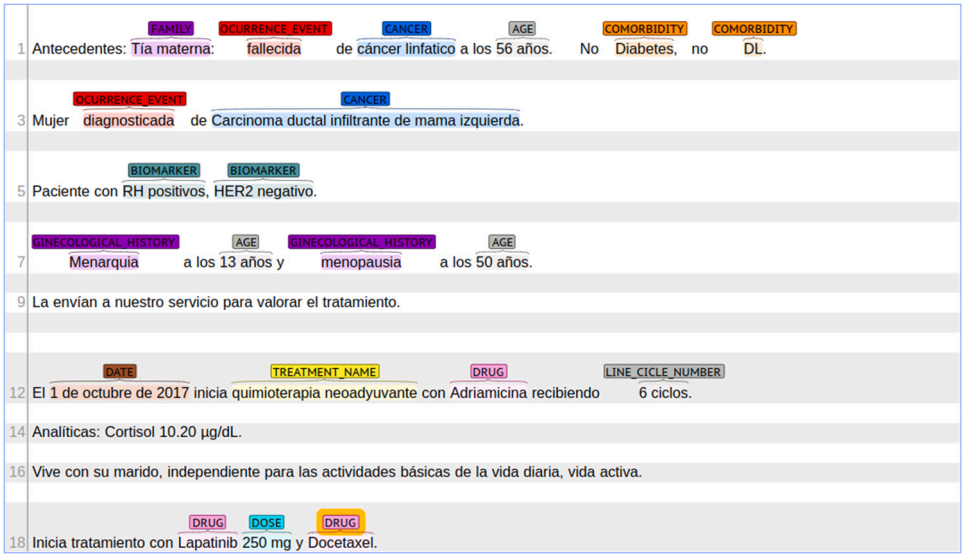

Fig. 2. Breast cancer annotations using the BRAT annotation tool.

- **Cancer stage** is used to identify the tumor stage when the patient was diagnosed. Staging is the process of finding out how far along the tumor is in the body and how far it has spread. The cancer stage can be described in two ways, as described in [68]:
  - Using a scale that ranges from I to IV, where "Stage I" represents the initial stage and "Stage IV" indicates the most advanced state of the disease.
  - Using the TNM notation that represents the following values: Tumor size (T), Nearby lymph nodes (N), and Metastasis (M). The next sentence uses this notation to represent the tumor stage: "Patient diagnosed with breast carcinoma, *cT3cN1cM1*".
- **Gynecological history** is used to represent events related to the female reproductive history. It includes concepts such as menarche or menopause.
- **Obstetrics history:** describes information such as number and type pregnancies, number of births and related.
- **Age:** is used for annotating the patient's age which physicians frequently associate with events in the patient history.
- **Date** is used to identify time expressions presented in clinical notes. Dates are frequently mentioned in clinical texts and describe the time when an event has occurred in the patient history, such as diagnosis, treatments, gynecological, obstetrics events, and others. Dates can be defined explicitly (e.g., "25/03/2007") or implicitly (e.g., "A month ago").
- **Therapy** identifies the name of different therapies used for treating cancer patients, such as chemotherapy, radiotherapy, immunotherapy, hormonal therapy, and others.
- **Medication** describes drug names used for treating cancer patients. In addition, we used the *Dose* label to represent the medication dosage.
- **Duration** is used to indicate a period of time during which the patient is undergoing treatment or have used a certain medication such as contraceptives.
- **Treatment line** is used to identify the number of cycles within a therapy applied to cancer patients.
- **Tumor progression** indicates a increased growth speed and invasiveness of the tumor cells. As a result of the progression, phenotypical changes occur and the tumor becomes more aggressive and acquires greater malignant potential.
- **Cancer biomarker:** is used for annotating different biomarkers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Extracting the values of these biomarkers are crucial to determine how a person diagnosed with breast cancer will be treated by clinicians.
- **Toxic habit** represents the toxic habits of patients, such as being a smoker or alcohol drinker. We also used the label Quantity Habits for representing a the number of units consumed. (e.g., the number of cigarettes the patient smoked.)
- **Allergy** describes allergies to medications or other substances that the patient can suffer.
- **Comorbidity** is used for representing another diseases or conditions that co-occurs with a cancer diagnosis (e.g., diabetes mellitus, high blood pressure)

In addition, we extend the above annotation schema with the entity types Surgery, Frequency, Patient events and Family members described in detail in [68].

3.1.2. *Annotating clinical notes*

Following the annotation schema previously described, the clinical notes are annotated. These annotations were created by two clinicians experts in the oncology domain. One data scientist who led the annotation process guided these annotators, prepared documents, and explained the annotation tool functionalities. The annotation process was developed in four steps: Anonymization, Annotation, Review, and Disagreement resolution.

- **Anonymization:** Clinical notes first were anonymized by eliminating patients or physicians data.
- **Annotation:** Clinicians independently annotate 100% of the clinical documents and they perform the process separately.
- **Review:** Annotators perform a review of all the annotations. They independently check all documents which they have annotated in the previous step and make modifications if necessary.
  The time to complete the annotation process took seven months.
- **Disagreement resolution**: The annotators were assisted by a data scientist who reviewed cases where there were disagreements and worked with clinicians to solve them.

3.1.3. *Measuring annotations reliability*

The annotations reliability was performed by calculating the inter-annotator agreement (IAA) for manually annotated corpora. The IAA

Table 2
Entity types and their corresponding number of annotations, along with their inter-annotation agreement (IAA).

Entity type	Number of annotations	IAA
Age	1184	1.00
Allergies	357	0.97
Biomarker	1855	0.96
Cancer entity	2535	0.91
Clinical service	538	0.95
Comorbidity	1385	0.89
Cycle number	1134	0.93
Date	2913	0.96
Duration	443	0.90
Dose	859	0.95
Drug	3233	0.94
Family	393	0.98
Frequency	960	0.92
Gynecological history	1034	0.95
Habits quantity	240	0.97
Implicit date	727	0.90
Obstetrics history	681	0.96
Occurrence event	2085	0.90
Progression	439	0.91
Surgery	1324	0.92
Toxic habits	476	0.95
Tumor stage	632	0.93
Tumor TNM	851	0.95

was obtained by measuring the F-measure between two annotators, which has been widely used for measuring the reliability in named entity annotation corpora [57,69–73]. @The IAA measure aims to guarantee that annotators produce similar and consistent annotations during the annotation process. The IAA was calculated at the end of the fourth step (Review) in the annotation process.

Similarly, as described in [74], we calculate the IAA as follows:

- Let A_1 be the first annotator, and let A_2 be the second annotator
- We take one set of annotations (A_1 or A_2) as the gold standard. Correct annotations are those made by this annotator. Consequently, Precision is the percentage of *correct positive* annotations produced by the other annotator. Recall is the percentage of *positive* annotations produced by the other annotator.
- F-measure is the harmonic mean between Precision and Recall, as shown in Eq. (1).

$$\textbf{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Table 2 shows the IAA using the F-measure, which was obtained at entity level, as follows:

- **Entity level:** the IAA is calculated at entity level. Correct annotations are those entities where both annotators agree on all annotated tokens. In the sentence shown above, an annotation is considered an agreement when the two annotators tagged the same tokens *"carcinoma ductal infiltrante"*. Otherwise, it is considered a disagreement.

3.2. Model training

In this approach clinical NER is performed as a sequence-labeling task, in which each entity in a text sentence is classified according to a specific label defined in the annotated corpus. We explore two transformer-based architectures to perform clinical NER in the cancer domain: BERT [33] and RoBERTa [34]. These artificial neural models have become common deep learning architectures to perform text sequence labeling tasks. The model training has been performed in two steps: Corpus Preprocessing, and Fine-tuning.

1. **Preprocessing**: This step transforms annotations in the corpus into the BIO tagging format where each annotated token is labeled with B (at the beginning of the entity), I (inside the entity), or O (Outside the entity).
   For instance, the sentence:
   "Mujer con carcinoma ductal infiltrante de mama". (The cancer concept is underlined) can be formatted as follows:
   {'O', 'O', 'B-CANCER', 'I-CANCER', 'I-CANCER', 'I-CANCER', 'I-CANCER', 'O'}
2. **Fine-tuning:** We perform a fine-tuning process using four different monolingual and multilingual models pre-trained for BERT and RoBERTa. The main characteristics of these models are shown in Table 3. Fig. 3 shows the fine-tuning process to perform clinical NER using the Breast cancer corpus. This process consists of three steps: Sentence tokenization, Transformer Processing, and Classification & Post-processing.

   - **Sentence tokenization:** the goal of this step is to tokenize a text sentence using a WordPiece Tokenization method [75]. For each word in the sentence, this method decides to keep the whole word or to split it into a set of sub-words. Additionally, in this step, two special tokens are added to the sentence: [CLS] and [SEP]. The [CLS] token always appears at the beginning of the text, and the [SEP] token is used to separate sentences.
     Fig. 3 shows an example in which the words *"en", "mayo", "de", "2018"* have not been split into sub-words because the WordPiece tokenization method considers these words as frequent words in its vocabulary. In contrast, the words *"carcinoma", "infiltrante"* are considered rare words by the algorithm as they do not belong to its vocabulary. Consequently, these words have been split in sub-words and the characters "##" are used to separate these tokens. The WordPiece Tokenization algorithm uses a Byte Pair Encoding strategy [76] to generate new sub-words.
   - **Transformer Processing:** in this step, the approach takes as input a tokenized sentence from the previous step and obtains an embedding representation (E_i) for each word in the sentence. This representation is created using three embeddings: token, segment, and position embeddings. Then, the Transformer Block takes the embedding representation as input (E_1, E_2, E_n) and produces a final representation (R_i) for each token in the processed sentence. This representation is a score calculated by the transformer and represents a contextualized value for a specific word in relation to all other words in the sentence.
   - **Classification & Post-Processing:** each predicted representation (R_i) is taken as input and fed into a dense layer with a softmax activation function. This layer obtains label for each token in the sentence, calculating a probability P for each label using the softmax function, as follows:

   $$P(l|\mathbf{R}_i) = \textbf{Softmax}(\mathbf{W}_o\mathbf{R}_i + \mathbf{b}_o) \quad (2)$$

   where the label l belongs to the set of labels to be predicted. In addition, W_o is a matrix of weight parameters and b_o is a bias vector, both learned by the dense layer. Finally, the special tokens "[CLS], [SEP], [PAD]" are removed to obtain the final BIO labels at the end of post-processing step.

3.3. Validation

In what follows we will describe the evaluation methodology to validate the presented approach as well as the hyperparameters configuration. We have validated the approach using several language models based on BERT and RoBERTa-based, as it is shown in Table 3.

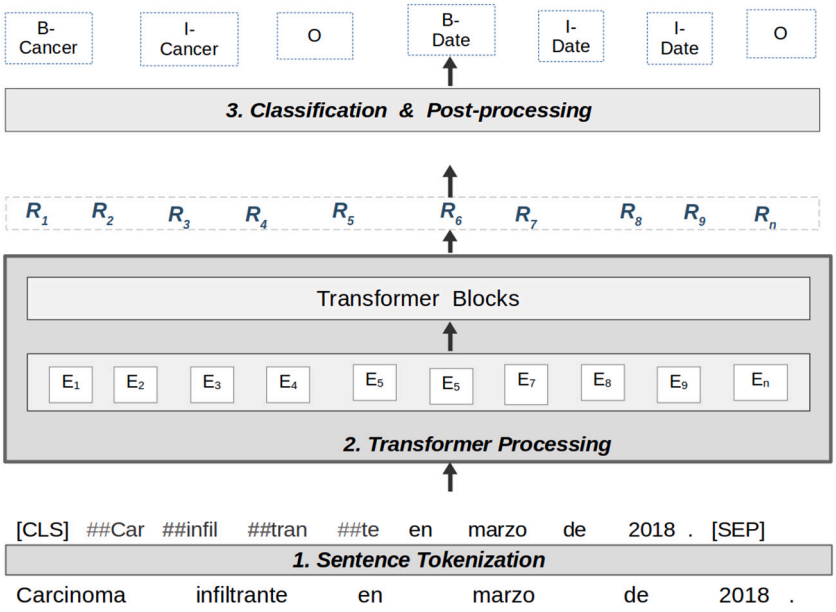

Fig. 3. Breast cancer information extraction using transformers.

Table 3
Pre-trained language models.

Transformer model	Architecture	Language	Vocabulary size	Text resources
BETO [38]	Bert-base	Spanish	31,000	Wikipedia, Nations and Government journals, TED Talks
Multilingual BERT [33]	Bert-base	Multilingual	120,000	Wikipedia
RoBERTa Biomedical [39]	Roberta-base	Spanish	38,000	Biomedical texts and clinical narratives
RoBERTa BNE [40]	Roberta-base	Spanish	50,262	National Library of Spain.

3.4. Evaluation methodology

To evaluate the performance of the proposed approach, we used the traditional standard metrics: Precision (P), Recall (R), and F-score (F1). The F-score is calculated as a weighted average of the Precision and Recall measurements. We used a 10-fold cross validation for training and evaluating clinical NER models. The performance was calculated as the average of all ten folds executed by the cross-validation strategy. The breast cancer corpus described in Section 3.1 has been used to train models.

We report results using a **strict match** criteria which measures the performance at entity level. According to this evaluation criteria, an entity is considered a correct prediction when the tokens extracted by the system *exact match* the tokens annotated by experts, and they have the same label annotated in the corpus.

$$\textbf{Precision} = \frac{\text{Number of entities correctly predicted}}{\text{Number of predicted entities}} \tag{3}$$

$$\textbf{Recall} = \frac{\text{Number of entities correctly predicted}}{\text{Number of entities in the dataset}} \tag{4}$$

$$\textbf{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

4. Results

In this section, we first explain the results obtained for each trained model with the breast cancer corpus, and later, we present the performance for each entity type. Models have been trained using spaCy command line interface.[10] The configuration hyperparameters are the same to all models and they are show in Table 4.

[10] https://spacy.io.

Table 4
Hyperparameters used for training the models.

Hyperparameter	Value	Hyperparameter	Value
Seed	0	Batch size buffer	256
Accumulate gradients	3	Batch size discard oversize	True
Dropout	0.1	Learn rate	Warmup-linear
Optimizer	Adam	Initial rate	0.00005
GPU allocator	Pytorch	Total steps	20 000
Batch size	2000	Warmup steps	250

Table 5
Global results for each model.

Model	Precision	Recall	F-score
BETO	0.9321 ± 0.0043	0.9414 ± 0.0013	0.9371 ± 0.0022
Multilingual BERT	0.9443 ± 0.0011	0.9478 ± 0.0007	0.9463 ± 0.0007
RoBERTa Biomedical	0.9459 ± 0.0007	0.9542 ± 0.0004	0.9501 ± 0.0005
RoBERTa BNE	0.9465 ± 0.0015	$0.9443 \pm 0,0010$	0.9454 ± 0.0011

Table 5 shows global results obtained for each model. We have performed ten iterations of 10-fold cross-validation in order to calculate statistical significance. These results represent the average over all entity types in the breast cancer corpus in the ten executions. All transformer-based models presented in Table 5 show an F-Score above 93%, which suggests that this approach is suitable for extracting named entities from breast cancer clinical notes written in Spanish. The best performance was obtained by the RoBERTa Biomedical model, which obtained an F-score of 0.9501. Although the vocabulary size of "RoBERTa Biomedical" is smaller than "Roberta BNE", the former performed better. This fact indicates that pre-trained models with biomedical texts perform better in extracting named entities in the cancer domain than using general domain models.

To calculate the statistical significance, the Friedman test in combination with the Bonferroni test, has been applied. The Friedman test is

Table 6
P-values of the BETO, Multilingual BERT, RoBERTa Biomedical, and RoBERTa BNE models for Precision.

	BETO	M. BERT	RoBERTa Bio.	RoBERTa BNE
BETO	1	0.036460	0.025283	0.039193
M. BERT	0.036460	1	0.000005	0.000008
RoBERTa Bio.	0.025283	0.000005	1	1
RoBERTa BNE	0.039193	0.000008	1	1

Table 7
P-values of the BETO, Multilingual BERT, RoBERTa Biomedical, and RoBERTa BNE models for Recall.

	BETO	M. BERT	RoBERTa Bio.	RoBERTa BNE
BETO	1	4.40e−14	8.68e−07	3.22e−07
M. BERT	4.40e−14	1	9.93e−20	6.40e−06
RoBERTa Bio.	8.68e−07	9.93e−20	1	9.89e−15
RoBERTa BNE	3.22e−07	6.40e−06	9.89e−15	1

Table 8
P-values of the BETO, Multilingual BERT, RoBERTa Biomedical, and RoBERTa BNE models for F-score.

	BETO	M. BERT	RoBERTa Bio.	RoBERTa BNE
BETO	1	8.21e−10	2.25e−06	5.25e−02
M. BERT	8.21e−10	1	2.37e−16	3.06e−06
RoBERTa Bio.	2.25e−06	2.37e−16	1	6.18e−10
RoBERTa BNE	5.25e−02	3.06e−06	6.18e−10	1

a non-parametric test that compares differences between two or more related samples in order to determine if there are significant differences between the samples. When the result of the Friedman test is significant ($p - value < 0.05$), a post-hoc method is used to determine the samples that are significantly different. In the context of the Friedman test, Bonferroni is used to compare each pair of samples and determine if there is a significant difference. In the multiple comparison matrices of the Bonferroni test, the corrected p-values smaller than the selected significance level ($p - value < 0.05$) indicate a significant difference between the corresponding samples.

When performing these two tests in our case, the Friedman test gave a p-value of 0.0001867 ($p - value < 0.05$), indicating a statistically significant difference between the groups being compared. @Consequently, the Bonferroni test was performed to identify which pairs of samples were significantly different for BETO [38], multilingual BERT, RoBERTa Biomedical [39] and RoBERTa BNE [40]. Tables for Precision (Table 6), Recall (Table 7) and F-score (Table 8) represent the results of pairwise comparisons. Each element in the table represents the p-value of the comparison of the involved groups.

The Bonferroni $p - values$ for Precision indicate statistical significance for all pairs, except for RoBERTa Biomedical and RoBERTa BNE. Similarly, the comparison tables for Recall and F-score also show statistical significance for all pairs. These values, together with the overall results from Table 5, show that RoBERTa Biomedical is the best model for extracting breast cancer-named entities in Spanish.

The high performance of RoBERTa Biomedical can be attributed to the fact it has been pre-trained with biomedical texts and clinical narratives written in Spanish. Although this performance is slightly higher than the one obtained with other models, their vocabulary size is smaller than other models, such as multilingual BERT (See Table 3). This suggests that in order to obtain a model with higher performance, pre-training with clinical texts in Spanish can be better than training with a larger vocabulary.

Regarding the performance obtained for each entity, Table 9 shows results for all entity types. The results presented in the table show the mean and standard deviation of the evaluation metrics (Precision, Recall, and F-score) obtained from the ten executions of the 10-fold cross-validation. One can observe that the values for F-score for all entities are consistently high, with values ranging from 0.8976 to 0.9858. The global mean F-score for all entities is 0.9519, indicating strong overall performance. This shows that the RoBERTa Biomedical model is performing well in identifying and classifying various types of medical entities. In particular, the model accurately finds entities such as Age, Drug, and Date with an F-score above 0.98. However, entities like Duration and Comorbidity have slightly lower F-score values due to the fact that identifying these entities is harder due to the variety of contexts in which they appear.

Table 9
Performance obtained for each entity using RoBERTa Biomedical model.

Entity type	Precision	Recall	F-score
Age	0.9818 ± 0.0037	0.9901 ± 0.0022	0.9858 ± 0.0002
Allergies	0.9096 ± 0.0070	0.9382 ± 0.0104	0.9228 ± 0.0085
Biomarker	0.9660 ± 0.0044	0.9698 ± 0.0029	0.9679 ± 0.0033
Cancer entity	0.9239 ± 0.0032	0.9289 ± 0.0028	0.9263 ± 0.0024
Clinical service	0.9226 ± 0.0138	0.9463 ± 0.0040	0.9335 ± 0.0076
Comorbidity	0.9065 ± 0.0054	0.9146 ± 0.0046	0.9103 ± 0.0045
Cycle number	0.9488 ± 0.0027	0.9586 ± 0.0047	0.9535 ± 0.0033
Date	0.9896 ± 0.0007	0.9916 ± 0.0015	0.9906 ± 0.0009
Dose	0.9641 ± 0.0053	0.9768 ± 0.0023	0.9703 ± 0.0030
Drug	0.9802 ± 0.0017	0.9821 ± 0.0020	0.9811 ± 0.0012
Duration	0.8879 ± 0.0145	0.9099 ± 0.0059	0.8976 ± 0.0088
Family	0.9674 ± 0.0054	0.9671 ± 0.0043	0.9671 ± 0.0036
Frequency	0.9366 ± 0.0045	0.9320 ± 0.0033	0.9340 ± 0.0030
Gynecological history	0.9375 ± 0.0075	0.9425 ± 0.0035	0.9398 ± 0.0046
Habits quantity	0.9520 ± 0.0192	0.9805 ± 0.0064	0.9654 ± 0.0131
Implicit date	0.9171 ± 0.0120	0.9507 ± 0.0044	0.9330 ± 0.0065
Obstetrics history	0.9407 ± 0.0059	0.9499 ± 0.0062	0.9450 ± 0.0051
Occurrence event	0.9092 ± 0.0038	0.9250 ± 0.0039	0.9168 ± 0.0021
Progression	0.9194 ± 0.0076	0.9313 ± 0.0046	0.9248 ± 0.0056
Surgery	0.9154 ± 0.0063	0.9292 ± 0.0046	0.9220 ± 0.0051
Toxic habits	0.9518 ± 0.0050	0.9592 ± 0.0071	0.9552 ± 0.0052
Tumor stage	0.9305 ± 0.0093	0.9412 ± 0.0066	0.9355 ± 0.0072
Tumor TNM	0.9565 ± 0.0061	0.9548 ± 0.0054	0.9555 ± 0.0054

Furthermore, a moderate correlation of 0.302 between the number of annotations per entity and the F-score has been observed. A notable finding is the strong correlation of 0.739 between the inter-annotation agreement (IAA) and the F-score, as shown in Fig. 4. This correlation is particularly interesting because it demonstrates that enhancing the agreement between annotators results in improved performance in the entity recognition task. Additionally, the correlation between the number of annotations and the F-score indicates that having more annotations per entity also leads to better performance. These findings show the importance of ensuring high-quality annotations in NLP tasks.

Thus, the presented approach demonstrates its ability to extract entities of varying lengths, entities with infrequent annotations, and entities with differing levels of variability in the breast cancer domain.

5. Discussion

In this paper, the process of annotating a breast cancer corpus has been shown. This corpus has been used to train transformer-based models to perform clinical NER. In particular, we have proposed an approach to perform clinical NER in the breast cancer domain from clinical narratives written in Spanish. This approach leverages transfer learning by using pre-trained language models to support clinical information extraction. Fine-tuning BERT [33] and RoBERTa [34] models have shown a feasible approach to extracting cancer information. As described in Table 9, this approach has shown promising results in extracting named entities in the breast cancer domain.

Annotating the corpus required a significant effort by oncology experts to create a reliable set of annotations. To measure the consistency of the corpus, the inter-annotator agreement between two annotators was calculated, obtaining values above 89% for each entity type. This means that annotations are consistent and reliable. Therefore, the presented corpus represents a valuable resource to support information extraction in the cancer domain.

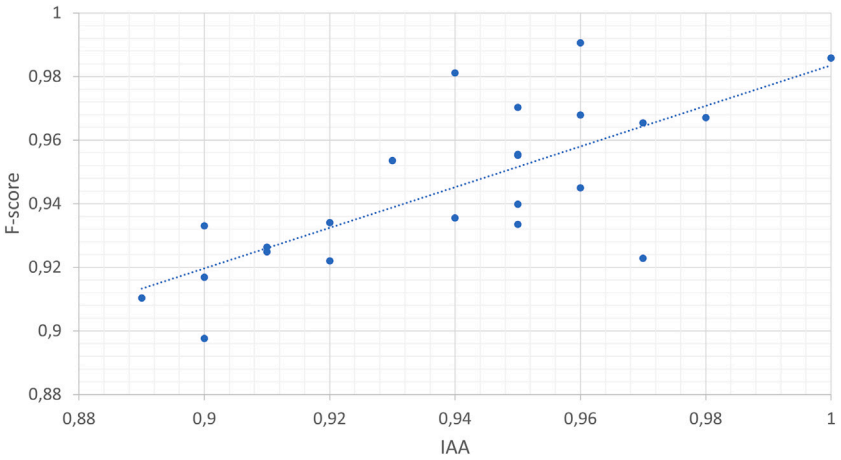

Fig. 4. Strong correlation between the inter-annotation agreement (IAA) and the F-score.

To support the extraction of the oncology-related entities, we have proposed an annotation schema to be used in the breast cancer domain, which has been shown in Section 3.1. The annotation schema includes twenty-three types ranging from antecedents and diagnoses to treatments. To our knowledge, this represents the first annotation schema for Spanish texts in the breast cancer domain. Besides, we have presented a transformer-based approach for performing clinical Named Entity Recognition (NER) in the breast cancer domain, focusing on clinical narratives written in Spanish.

The results indicate that the RoBERTa Biomedical model outperforms the other models, achieving the highest average F-score across all entity types. The superior performance of the RoBERTa Biomedical model can be attributed to its pre-training on biomedical texts and clinical narratives written in Spanish. This demonstrates the potential advantage of using domain-specific pre-trained models for clinical NER tasks, even when a smaller vocabulary size has been used compared to more general models, such as Multilingual BERT.

The performance for each entity type has shown promising results, with F-score values ranging from 0.8976 to 0.9858. Entities such as Age, Drug, and Date achieved particularly high F1 scores, while entities like Duration and Comorbidity showed slightly lower scores. This can be due to the varying complexity of the text and the context in which these entities appear.

Our analysis also revealed a strong correlation (0.739) between inter-annotation agreement (IAA) and the F-score, highlighting the importance of high-quality annotations in NLP tasks. This finding suggests that improving the agreement between annotators can lead to better performance in NER tasks. Additionally, a moderate correlation (0.302) between the number of annotations per entity and the F-score implies that having more annotations per entity can also contribute to better performance.

The study demonstrates that the transformer-based approach can effectively address challenges in extracting clinical entities in the cancer domain, such as dealing with a variety of semantic types, family history, entities with different lengths, low-frequency annotations, and varying levels of variability.

Results obtained with this research were used to structure the data of a hospital in Madrid to extract models for breast cancer relapse and analysis of quality of life. In [77], some results of the patient profiles are shown.

The proposed approach leveraging transfer learning with pre-trained language models shows great potential for extracting named entities in the breast cancer domain. The results underline the importance of domain-specific pre-training and high-quality annotations in achieving strong performance in clinical NER tasks. Future work could explore additional ways to improve the models, such as incorporating domain-specific knowledge or integrating other sources of information.

6. Conclusions

This study presents a comprehensive approach to performing clinical Named Entity Recognition (NER) in the breast cancer domain, specifically for clinical narratives written in Spanish. The approach is based on the use of transformer-based models, leveraging transfer learning through pre-trained language models such as BERT and RoBERTa. The breast cancer corpus was manually annotated, which required significant effort by oncology experts to ensure consistency and reliability in the annotations. The results show that the RoBERTa Biomedical model outperforms other models in terms of Precision, Recall, and F-score.

This research demonstrates the potential of transformer-based models in dealing with challenges related to extracting clinical entities in the cancer domain, such as diverse semantic types, family history, different entity lengths, and varying frequencies of annotations. The strong correlation between inter-annotation agreement (IAA) and F-score highlights the importance of high-quality annotations in NLP tasks.

The proposed annotation schema for the breast cancer domain represents a valuable resource for information extraction in the cancer domain, as it encompasses twenty-three entity types. The best results were achieved with the RoBERTa Biomedical model, despite its more limited vocabulary compared to other models. This suggests that pre-training with clinical texts in Spanish yields better performance than models with larger vocabularies.

Results have been applied in a real use case to structure the information of breast cancer patients of a hospital in Madrid, which is now being used to calculate risk factors, patterns for survivorship, and relapse models, among others.

Overall, the proposed approach demonstrates promising results in extracting named entities in the breast cancer domain from clinical narratives in Spanish, indicating the potential of transformer-based models to support clinical information extraction.

Declaration of competing interest

I confirm that there is not conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome. As Corresponding author, I confirm that the manuscript has been read and approved for submission by all the named authors.

Data availability

Data and code are available in: https://github.com/solarte7/cancerNERSpanish

The breast cancer corpus is available "upon request". This corpus can be accessible after an evaluation by the hospital's ethics committee. To request access to the anonymized data, please contact Dr. Maria Torrente at the following email: maria.torrente@salud.madrid.org

Acknowledgments

This paper is supported by European Union's Horizon2020 research and innovation programme under grant agreement number 875160 and by Fundacion AECC and Instituto de Salud Carlos III (grant AC19/00034), under the frame of ERA-NET PerMed. The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA: Cancer J Clin 2022.

[2] Kehl KL, Xu W, Lepisto E, Elmarakeby H, Hassett MJ, Van Allen EM, Johnson BE, Schrag D. Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clin Cancer Inform 2020;1(4):680–90. http://dx.doi.org/10.1200/CCI.20.00020, PMID: 32755459.

[3] Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. Appl Clin Inform 2018;9(01):046–53.

[4] Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Appl Sci (Switzerland) 2021;11(18). http://dx.doi.org/10.3390/app11188319.

[5] Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs): A survey. ACM Comput Surv 2018;50(6). http://dx.doi.org/10.1145/3127881.

[6] Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. In: AMIA ... annual symposium proceedings, vol. 2017. AMIA Symposium; 2017, p. 1812–9.

[7] Dalianis H. Clinical text mining. Springer Open; 2018, http://dx.doi.org/10.1007/978-3-319-78503-5.

[8] Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinform 2019;16(1):139–53. http://dx.doi.org/10.1109/TCBB.2018.2849968, arXiv:1806.04820.

[9] Zhou Y, Ju C, Caufield JH, Shih K, Chen C, Sun Y, Chang K-W, Ping P, Wang W. Clinical named entity recognition using contextualized token representations. 2021, Arxiv. arXiv:2106.12608. URL http://arxiv.org/abs/2106.12608.

[10] Yang X, Mu D, Peng H, Li H, Wang Y, Wang P, Wang Y, Han S, et al. Research and application of artificial intelligence based on electronic health records of patients with cancer: Systematic review. JMIR Med Inform 2022;10(4):e33799.

[11] Chen X, Liu Z, Wei L, Yan J, Hao T, Ding R. A comparative quantitative study of utilizing artificial intelligence on electronic health records in the USA and China during 2008–2017. BMC Med Inform Decis Making 2018;18(5):55–69.

[12] Yim WW, Yetisgen M, Harris WP, Sharon WK. Natural language processing in oncology review. JAMA Oncol 2016;2(6):797–804. http://dx.doi.org/10.1001/jamaoncol.2016.0213.

[13] Keretna S, Lim CP, Creighton D, Shaban KB. Enhancing medical named entity recognition with an extended segment representation technique. Comput Methods Programs Biomed 2015;119(2):88–100. http://dx.doi.org/10.1016/j.cmpb.2015.02.007.

[14] Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, Shen F, Wang L, Wang Y, Wen A, Zhao Y, Sohn S, Liu H. Clinical concept extraction: A methodology review. J Biomed Inform 2020;109:103526. http://dx.doi.org/10.1016/j.jbi.2020.103526.

[15] Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: Challenges and opportunities. In: 2016 IEEE international conference on big data (big data). 2016, p. 1937–45. http://dx.doi.org/10.1109/BigData.2016.7840814.

[16] Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010;17(4):440–5. http://dx.doi.org/10.1136/jamia.2010.003707.

[17] Evans TL, Gabriel PE, Shulman LN. Cancer staging in electronic health records: Strategies to improve documentation of these critical data. J Oncol Pract 2016;12(2):137–9. http://dx.doi.org/10.1200/jop.2015.007310.

[18] Khor RC, Nguyen A, O'Dwyer J, Kothari G, Sia J, Chang D, Ng SP, Duchesne GM, Foroudi F. Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. Int J Med Inform 2019;121(April 2017):53–7. http://dx.doi.org/10.1016/j.ijmedinf.2018.10.008.

[19] Zhou H, Ning S, Yang Y, Liu Z, Xu J. Chinese hedge scope detection based on phrase semantic representation. In: 2017 International conference on asian language processing, vol. 2018-janua. IALP, 2018, p. 285–8. http://dx.doi.org/10.1109/IALP.2017.8300599.

[20] Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS ONE 2012;7(1). http://dx.doi.org/10.1371/journal.pone.0030412.

[21] Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, Tulsky JA, Lindvall C. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage 2018;55(6):1492–9. http://dx.doi.org/10.1016/j.jpainsymman.2018.02.016.

[22] Isaksson LJ, Pepa M, Zaffaroni M, Marvaso G, Alterio D, Volpe S, Corrao G, Augugliaro M, Starzyńska A, Leonardi MC, Orecchia R, Jereczek-Fossa BA. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. Front Oncol 2020;10(June). http://dx.doi.org/10.3389/fonc.2020.00790.

[23] Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning based text classification: A comprehensive review, arXiv 2020;1(1)1–43. arXiv:2004.03705.

[24] Wang L, Luo L, Wang Y, Wampfler J, Yang P, Liu H. Natural language processing for populating lung cancer clinical research data. BMC Med Inform Decis Mak 2019;19(Suppl 5):1–10. http://dx.doi.org/10.1186/s12911-019-0931-8.

[25] Li L, Zhao J, Hou L, Zhai Y, Shi J, Cui F. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. BMC Med Inform Decis Mak 2019;19(Suppl 5):1–11. http://dx.doi.org/10.1186/s12911-019-0933-6.

[26] Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics 2019;20(1):1–11. http://dx.doi.org/10.1186/s12859-019-3321-4.

[27] Wang Y, Sun Y, Ma Z, Gao L, Xu Y. Named entity recognition in Chinese medical literature using pretraining models. Sci Program 2020;2020. http://dx.doi.org/10.1155/2020/8812754.

[28] Bitterman D, Miller T, Harris D, Lin C, Finan S, Warner J, Mak R, Savova G. Extracting radiotherapy treatment details using neural network-based natural language processing. Int J Radiat Oncol, Biol, Phys 2020;108(3):e771–2.

[29] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than english: Opportunities and challenges. J Biomed Semant 2018;9(1):1–13. http://dx.doi.org/10.1186/s13326-018-0179-8.

[30] Pabón OS, Montenegro O, Torrente M, González AR, Provencio M, Menasalvas E. Negation and uncertainty detection in clinical texts written in spanish: a deep learning-based approach. PeerJ Comput Sci 2022;8:e913.

[31] Miranda-Escalada A, Farré E, Krallinger M. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In: CEUR workshop proceedings, vol. 2664. 2020, p. 303–23.

[32] Najafabadipour M, Tuñas JM, Rodríguez-González A, Menasalvas E. Lung cancer concept annotation from spanish clinical narratives. In: Auer S, Vidal M-E, editors. Data integration in the life sciences. Springer International Publishing; 2019, p. 153–63.

[33] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies - proceedings of the conference, vol. 1. 2019, p. 4171–86, (Mlm) arXiv:1810.04805.

[34] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. 2019, arXiv preprint arXiv:1907.11692.

[35] Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, Bian J, He Z. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics. BCB '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 1–6. http://dx.doi.org/10.1145/3459930.3469560.

[36] Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, Armentano-Oller C, Rodriguez-Penagos C, Gonzalez-Agirre A, Villegas M. Maria: Spanish language models. Proces Del Leng Nat 2022;68:39–60.

[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International conference on neural information processing systems. NIPS '17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 6000–10.

[38] Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J. Spanish pre-trained BERT model and evaluation data. In: PML4DC at ICLR 2020. 2020, p. 1–6.

[39] Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, Armengol-Estapé J, Silveira-Ocampo J, Valencia A, Gonzalez-Agirre A, Villegas M. Pretrained biomedical language models for clinical NLP in spanish. In: Proceedings of the 21st workshop on biomedical language processing. Dublin, Ireland: Association for Computational Linguistics; 2022, p. 193–9. http://dx.doi.org/10.18653/v1/2022.bionlp-1.19, URL https://aclanthology.org/2022.bionlp-1.19.

[40] Gutiérrez Fandiño A, Armengol Estapé J, Pàmies M, Llop Palao J, Silveira Ocampo J, Pio Carrino C, Armentano Oller C, Rodriguez Penagos C, Gonzalez Agirre A, Villegas M. Maria: Spanish language models. Proces Del Leng Nat 2022;68.

[41] Tulkens S, Šuster S, Daelemans W. Unsupervised concept extraction from clinical text through semantic composition. J Biomed Inform 2019;91(February):103120. http://dx.doi.org/10.1016/j.jbi.2019.103120.

[42] Saiz FS, Sanders C, Stevens R, Nielsen R, Britt M, Yuravlivker L, Preininger AM, Jackson GP. Artificial intelligence clinical evidence engine for automatic identification, prioritization, and extraction of relevant clinical oncology research. JCO Clin Cancer Inform 2021;5:102–11.

[43] Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. Recap: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract 2016;12(2):157–8. http://dx.doi.org/10.1200/jop.2015.004622.

[44] Soysal E, Warner JL, Wang J, Jiang M, Harvey K, Jain SK, Dong X, Song H-Y, Siddhanamatha H, Wang L, et al. Developing customizable cancer information extraction modules for pathology reports using clamp. Stud Health Technol Inform 2019;264:1041–5. http://dx.doi.org/10.3233/SHTI190383.

[45] Liu K, Mitchell KJ, Chapman WW, Crowley RS. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. In: AMIA ... Annual symposium proceedings / AMIA symposium, vol. 11. AMIA Symposium; 2005, p. 460–4, (Figure 1).

[46] Dienstmann R, Mason MJ, Sinicrope FA, Phipps AI, Tejpar S, Nesbakken A, Danielsen SA, Sveen A, Buchanan DD, Clendenning M, Rosty C, Bot B, Alberts SR, Milburn Jessup J, Lothe RA, Delorenzi M, Newcomb PA, Sargent D, Guinney J. Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. Ann Oncol : Off J Eur Societ Med Oncol 2017;28(5):1023–31. http://dx.doi.org/10.1093/annonc/mdx052.

[47] Yim W-w, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform 2016;64:179–91.

[48] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory. 1992, p. 144–52.

[49] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning. ICML '01, Williamstown, MA, USA: Morgan Kaufmann Publishers Inc.; 2001, p. 282–9, URL http://dl.acm.org/citation.cfm?id=645530.655813.

[50] Weegar R, Kvist M, Sundström K, Brunak S, Dalianis H. Finding cervical cancer symptoms in Swedish clinical text using a machine learning approach and negex. In: AMIA annual symposium proceedings, vol. 2015. American Medical Informatics Association; 2015, p. 1296.

[51] Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, Mani S, Levy MA, Dai Q, Denny JC. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. In: AMIA annual symposium proceedings, vol. 2011. American Medical Informatics Association; 2011, p. 1564.

[52] Lenain R, Seneviratne MG, Bozkurt S, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine learning approaches for extracting stage from pathology reports in prostate cancer. Stud Health Technol Inform 2019;264:1522.

[53] Oliwa T, Maron SB, Chase LM, Lomnicki S, Catenacci DV, Furner B, Volchenboum SL. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. JCO Clin Cancer Inform 2019;3(3):1–8. http://dx.doi.org/10.1200/cci.19.00008.

[54] Martina S, Ventura L, Frasconi P. Classification of cancer pathology reports: A large-scale comparative study. IEEE J Biomed Health Inf 2020;24(11):3085–94. http://dx.doi.org/10.1109/JBHI.2020.3005016, arXiv:2006.16370.

[55] Wang Y, Sun Y, Ma Z, Gao L, Xu Y. Named entity recognition in Chinese medical literature using pretraining models. Sci Program 2020;2020. http://dx.doi.org/10.1155/2020/8812754.

[56] Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019;132(September):103985. http://dx.doi.org/10.1016/j.ijmedinf.2019.103985.

[57] Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I, Rüter G, Hautow H, Sänger M, Habibi M, Zettwitz M, de Bortoli T, Ostermann L, Ševa J, Starlinger J, Kohlbacher O, Malek NP, Keilholz U, Leser U. Annotation and initial evaluation of a large annotated german oncological corpus. JAMIA Open 2021;4(2):1–9. http://dx.doi.org/10.1093/jamiaopen/ooab025.

[58] Garciá-Pablos A, Perez N, Cuadros M. Vicomtech at cantemist 2020. In: CEUR workshop proceedings, vol. 2664. 2020, p. 489–98.

[59] Carrasco SS, Martínez P. Using embeddings and bi-lstm+crf model to detect tumor morphology entities in spanish clinical cases. In: CEUR workshop proceedings, vol. 2664. 2020, p. 368–75.

[60] López-Úbeda P, Diáz-Galiano MC, Martín-Valdivia MT, Urenã-López LA. Extracting neoplasms morphology mentions in spanish clinical cases throughword embeddings. In: CEUR workshop proceedings, vol. 2664. 2020, p. 324–34.

[61] Bhatt AN, Mathur R, Farooque A, Verma A, Dwarakanath B. Cancer biomarkers-current perspectives. Indian J Med Res 2010;132(2):129–49.

[62] Morgado J, Pereira T, Silva F, Freitas C, Negrão E, de Lima BF, da Silva MC, Madureira AJ, Ramos I, Hespanhol V, Costa JL, Cunha A, Oliveira HP. Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. Appl Sci (Switzerland) 2021;11(7). http://dx.doi.org/10.3390/app11073273.

[63] Zolbanin HM, Delen D, Hassan Zadeh A. Predicting overall survivability in comorbidity of cancers: A data mining approach. Decis Support Syst 2015;74:150–61. http://dx.doi.org/10.1016/j.dss.2015.04.003.

[64] Nguyen G, Dlugolinsky S, Tran V, Lopez Garcia A. Deep learning for proactive network monitoring and security protection. IEEE Access 2020;8:19696–716. http://dx.doi.org/10.1109/ACCESS.2020.2968718.

[65] Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, Tulsky JA, Lindvall C. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage 2018;55(6):1492–9. http://dx.doi.org/10.1016/j.jpainsymman.2018.02.016.

[66] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. 2015, arXiv:1508.01991, URL http://arxiv.org/abs/1508.01991.

[67] Bitterman D, Miller T, Harris D, Lin C, Finan S, Warner J, Mak R, Savova G. Extracting relations between radiotherapy treatment details. In: Proceedings of the 3rd clinical natural language processing workshop. Online: Association for Computational Linguistics; 2020, p. 194–200. http://dx.doi.org/10.18653/v1/2020.clinicalnlp-1.21, URL https://www.aclweb.org/anthology/2020.clinicalnlp-1.21.

[68] Solarte-Pabón O, Blazquez-Herranz A, Torrente M, Rodríguez-Gonzalez A, Provencio M, Menasalvas E. Extracting cancer treatments from clinical text written in spanish: A deep learning approach. In: 2021 IEEE 8th international conference on data science and advanced analytics. DSAA, 2021, p. 1–6. http://dx.doi.org/10.1109/DSAA53316.2021.9564137.

[69] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296–8.

[70] Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc 2015;22(1):143–54. http://dx.doi.org/10.1136/amiajnl-2013-002544.

[71] Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. J Biomed Inform 2015;56:318–32. http://dx.doi.org/10.1016/j.jbi.2015.06.016.

[72] Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A, Moreno-Sandoval A. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. BMC Med Inform Decis Mak 2021;21(1):1–19. http://dx.doi.org/10.1186/s12911-021-01395-z.

[73] Savkov A, Carroll J, Koeling R, Cassell J. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. Lang Resour Eval 2016;50(3):523–48. http://dx.doi.org/10.1007/s10579-015-9330-7.

[74] Alnazzawi N, Thompson P, Ananiadou S. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In: Proceedings of the 5th international workshop on health text mining and information analysis (Louhi). Gothenburg, Sweden: Association for Computational Linguistics; 2014, p. 69–74. http://dx.doi.org/10.3115/v1/W14-1110, URL https://aclanthology.org/W14-1110.

[75] Song X, Salcianu A, Song Y, Dopson D, Zhou D. Fast WordPiece tokenization. 2020, arXiv preprint arXiv:2012.15524.

[76] Schuster M, Nakajima K. Wordpiece tokenization. In: ICASSP, IEEE International conference on acoustics, speech and signal processing - proceedings, vol. 1. 2012, p. 5149–52.

[77] Torrente M, Sousa PA, Hernández R, Blanco M, Calvo V, Collazo A, Guerreiro GR, Núñez B, Pimentao J, Sánchez JC, et al. An artificial intelligence-based tool for data analysis and prognosis in cancer patients: Results from the clarify study. Cancers 2022;14(16):4041.