Artificial Intelligence for Natural Language Processing of Clinical Text in Spanish for Real-World-Data Analysis (Text2RWD Project)

Fancisco J. Veredas[1,2], Fernando Gallego[1,2], Guillermo López-García[1,2], Nuria Ribelles[3], Emilio Alba[3] and José M. Jerez[1,2]

[1]Grupo de Inteligencia Computacional en Biomedicina (ICB), Dept. Lenguajes y Ciencias de la Computación, Universidad de Málaga, 29071, Málaga, Spain

[2]Research Institute of Multilingual Language Technologies. Universidad de Málaga, 29071, Málaga, Spain

[3]Unidad de Gestión Clínica Intercentros de Oncología (UGCIO), Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain

Abstract

The Text2RWD project (funded by the Spanish Ministerio de Ciencia e Innovación, PID2020-116898RB-I00), aims to advance in the creation and de-identification of a specific clinical corpus, which is expected to be of reference as an oncological text corpus in Spanish. By using this corpus, new artificial intelligence (AI) algorithms for natural (Spanish-)language processing (NLP) are being designed and adapted to be applied to information-processing downstream tasks carried out on unstructured textual data stored in oncological electronic health records (EHR) contained in Galén, a healthcare information management system. The resulting models are to be analyzed and validated by applying them to the resolution of different clinically-significant tasks through the analysis of real world data in oncology units. The AI-for-NLP models are also expected to be transferred and applied to text corpora of other medical disciplines or healthcare settings, and validated in tackling information extraction and prediction tasks in those specific areas.

Keywords

natural language processing, artificial intelligence, electronic health records, real-world data, oncology, biomedical applications

1. Introduction and Background

The Text2RWD project, funded by the Spanish Ministerio de Ciencia e Innovación (reference number PID2020-116898RB-I00), is expected to last four years (2021-2025) and represents the continuation of a consolidated trajectory of our research team in the last 15 years, funded in successive calls by the Plan Nacional de I+D+i and by the Junta de Andalucía, through the execution of different projects in which progress has been made in the design of neuro-computational algorithms for cancer survival analysis (TIN2005-02984), constructive algorithms of artificial neural networks for information processing and data mining in oncology (TIN2008-04985, TIC-4026), bio-inspired intelligent systems applied to personalized medicine (TIN2010-16556), and the design and adaptation of deep learning (DL) algorithms for their application to problems in the field of biomedicine (TIN2017-88278). On the other hand, the FIMABIS Foundation (Andalusian Public Foundation for Research in Biomedicine and Health in Málaga) has funded, since 2015, the development of the new version of Galén, an information system that implements oncological electronic health records (EHR) and has been designed with the following three main objectives: i) to manage all the procedures that take place in a medical oncology unit; ii) to transfer the research results of the aforementioned projects to daily clinical practice; and iii) to provide an intelligent analysis of "real-world data" (RWD) or "real-life data" on cancer. This information system, which has been developed since 2008 by our research team in collaboration with the expert clinical staff of the *Unidad de Gestión Clínica Intercentros de Oncología* (UGCIO) of the *Hospital Universitario Virgen de la Victoria* (HUVV) and *Hospital Regional Universitario* (HRU) of Málaga, currently stores information relating to more than 60,000 cancer patients (1978 - 2023) in the province of Málaga, contains approximately 600,000 documents associated with the EHR and is integrated into Diraya, the information management system for healthcare in Andalusia.

The generalized use of EHR in the field of oncology constitutes an extremely valuable source of data, which due to its magnitude and veracity allow observational studies based on them to provide evidence that is hardly questionable. This type of information is known as

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

franveredas@uma.es (F. J. Veredas); fgdc2f3@uma.es (F. Gallego); guilopgar@uma.es (G. López-García); nuriaribelles@gmail.com (N. Ribelles); ealbac@uma.es (E. Alba); jmjerez@uma.es (J. M. Jerez)

https://sites.google.com/uma.es/franveredas (F. J. Veredas)

0000-0003-0739-2505 (F. J. Veredas); 0000-0001-5903-1483 (G. López-García); 0000-0003-3194-7398 (N. Ribelles); 0000-0002-3364-2603 (E. Alba); 0000-0002-7858-2966 (J. M. Jerez)

"real-world data" or "data from the real world", a concept that can be defined as all data related to the health status of patients or diagnostic and therapeutic procedures performed during daily clinical practice. Specifically, this concept refers to a huge volume of information of heterogeneous nature, in a wide variety of formats, which includes structured data of diverse clinical-pathological type and molecular expression data, as well as free-text documents (clinical notes on first encounters, subsequent encounters, emergency reports, pathological anatomy reports, etc.) that contain highly relevant information related to diagnoses, treatments and clinical procedures [1, 2, 3]. The extraction and exploitation of this unstructured information constitutes currently one of the main challenges for health systems in general and oncology settings in particular, insofar as it is considered of vital importance for the achievement of the main objective of advanced medicine: a more personalized, proactive, preventive and predictive healthcare. It is, indeed, an ambitious objective that raises a series of questions to the scientific community that remain partially unsolved, so that in recent years there has been a growing interest in the exploitation of clinical texts through the application of techniques and algorithms of natural language processing (NLP) and artificial intelligence (AI).

However, most of the previous studies of application of NLP and AI to RWD analysis problems, such as clinical coding or automatic document classification, are addressed on texts in English, due to the limited availability of annotated corpora with clinical coding information and additional linguistic resources in languages other than English. With about 470 million native speakers, there is growing interest in clinical text processing in Spanish. Proof of this are the "shared tasks" promoted by organizations such as CLEF eHealth Lab, for tasks such as automatic clinical coding on multilingual corpora or corpora in languages other than English [4, 5]. Specifically, the research team of this project has recently participated in the CANTEMIST task (CANcer TExt Mining Shared Task), promoted by the "Plan for the Promotion of Language Technologies" (Plan-TL) of the *Secretaría de Estado de Digitalización e Inteligencia Artificial* in Spain, for automatic labeling of oncological terms in clinical texts in Spanish related to the morphology of neoplasms (eCIE-O-3.1). The proposal of our research team [6], based on the use of the BERT attention model together with unstructured information stored in the Galén system, has been awarded the first prize of among 160 proposals from international research groups and companies.

In general, the design of algorithms for annotation and automatic classification of clinical texts requires manual retrospective analysis of a large amount of unstructured information, for the annotation and labeling of the categories corresponding to the documents used in the design of classification models. In this sense, the availability of the Galén system, with information on more than 60,000 cancer patients, with a total of ∼600,000 documents corresponding to clinical episodes and a significant number of structured fields (which are completed both in real time, during clinical care activity, and later by specific personnel in charge of this supervised task), supplies the research team with quality supervised information to address different types of classification and coding problems in the NLP field. Previously, and since 2003, our research team has been working actively in collaboration with the staff of the UGCIO in the design of predictive models of survival in cancer, with a large number of works published in journals both in the field of oncology and AI, most of them using structured clinical-pathological and molecular expression data extracted from Galén. Specifically, in the last years, our team has made significant progress in the design of predictive models for the evolution of patients with metastatic breast cancer, based exclusively on unstructured clinical text contained in EHRs of Galén, using text mining techniques in combination with AI algorithms [2]. These results, together with those also obtained in the design and adaptation of deep learning (DL) algorithms for their application to small datasets —using techniques such as transfer learning (TL) and data augmentation (DA) [7, 8]—, have provided this research team with the necessary experience to incorporating unstructured information into the design of classification and prediction models in the field of oncology, through the application of techniques and algorithms of AI to NLP. In fact, the incorporation of AI models to NLP tasks has significantly and surprisingly improved the efficiency of the classification and prediction algorithms applied in a wide variety of problems, such as conversational systems, automatic translation of texts, sentiment analysis in social networks and, of course, automatic document classification.

Thus, the Text2RWD is aiming to provide our research team with the possibility of advancing in i) the creation of a specific clinical de-identified corpus, expected to be of reference for the international of oncology in Spanish; ii) the design and adaptation of AI and NLP algorithms for their application to the processing of unstructured information contained in oncological EHRs in Spanish; iii) the application of the AI models to the resolution of different NLP downstream tasks in which the research team is currently working, in collaboration with UGCIO staff: detection/prediction of events (disease relapse, disease progression, patient death, prediction of emergency encounters, etc.), TNM classification of neoplasms and standardized ICD-10 coding; and iv) the transfer of the resulting models to the processing of EHR and text corpora in Spanish belonging to other medical disciplines, as well as to other healthcare settings.

2. Progress of the Text2RWD Project

In a phase prior to the development of the Text2RWD project, progress has been made in a line of research consisting of the application of TL approaches to problem solving in the biomedical field [9, 10, 7]. With this previous experience, the group has continued the same line of work by applying a TL-based strategy to address the problem of standardization of medical entities and automatic clinical coding of EHR texts in Spanish, using Transformer-based models. There is preliminary work by the ICB-UGCIO group itself in which BERT-based models have been applied with moderate results to specific clinical coding tasks in Spanish [11, 6]. However, the need for studies that systematically analyzed the performance of transformers in the Spanish medical setting, in particular for distinct clinical named-entity recognition problems, was noted. Moreover, there was also a lack of publicly available models based on transformers pre-trained on Spanish clinical corpora that could facilitate the adoption of these models in subsequent medical NLP tasks. All these reasons have justified the need to advance in this line of work, with the objectives and results presented below.

2.1. EHR De-identification

The primary objective in managing information within EHRs involves the automatic extraction and masking of concepts associated with personally identifiable data. This process, known as de-identification, is not only an ethical imperative but also a legal mandate stipulated by data privacy laws. Both the General Data Protection Regulation (GDPR) of the European Union (EU) and the *Ley Orgánica Española de Protección de Datos Personales y Garantía de Derechos Digitales* (LOPD-GDD) of Spain specifically prohibit the processing of personal data unless identifiable information is appropriately masked. As part of the Text2RWD project, we advanced in the development of AI and NLP algorithms for the de-identification of Spanish EHRs to ensure compliance with the LOPD-GDD.

For that purpose, in [8] we annotated a private corpus consisting of 599 real-world clinical cases with 8 distinct categories of protected health information. Addressing the predictive challenge as a named-entity recognition task, we developed two distinct methodologies rooted in DL. The first strategy employs recurrent neural networks (RNN), while the second adopts an end-to-end approach based on transformers. To augment the training data, we introduced a DA procedure, expanding the text corpus. Our findings indicate that transformers exhibit superior performance over RNN in the de-identification of Spanish clinical data. Notably, the XLM-RoBERTa-large Transformer demonstrated the best results, achieving a strict-match micro-averaged precision of 0.946, recall of 0.954, and an F1-score of 0.95 when trained on the augmented corpus [8]. The success of transformers in this study underscores their applicability in real-world clinical scenarios, showcasing the efficacy of these state-of-the-art models.

2.2. Oncology Text Classification

Addressing a recurring challenge within oncology clinical analysis units, the timely completion of first visit reports is hindered by the limited availability of clinical staff. Specifically, the structured fields, such as neoplasm type, location, and histology, often lack comprehensive information due to time constraints. Consequently, accessing and utilizing the results becomes exceedingly challenging. Although the necessary information exists within EHR, it is in an unstructured text format, dispersed throughout the EHR data rather than consolidated in dedicated electronic fields. The crucial task at hand involves automating the extraction of neoplasm type from the patient's EHR text, enabling the oncology analysis unit to promptly direct the patient to the appropriate specialist.

Within our Text2RWD project [12], we aim to advance the application of NLP models for the automated extraction of neoplasm types from EHRs written in Spanish. Our classification algorithms determine the likelihood of each document belonging to one of the three most prevalent neoplasms in the Galén information system: breast, colorectal, and lung; or being categorized as another type of neoplasm. The machine learning (ML) and DL models explored in this study included RNNs in conjunction with convolutional neural networks (CNN) and embedding models, which gave high performance rates in the classification task: 0.981 precision, 0.984 recall and 0.982 F1-score [12]. Notably, this study is the first of its kind, examining the application of NLP models to the task of extracting information about a patient's neoplasm from real-world medical texts in Spanish.

2.3. Clinical coding in Spanish

One of the most fundamental and challenging tasks in the medical NLP domain is automatic clinical coding. This task involves the conversion of unstructured clinical texts, written in specialized natural language, into structured formats that align with standardized coding terminologies, employing computational methods. In the Text2RWD, we primarily focus on the problem of automatic clinical coding for documents in Spanish. However, most of the existing literature has centered on English-written texts. This can be attributed to the scarce availability of corpora annotated with standardized clinical coding labels and supplementary linguistic resources in

languages other than English. Consequently, in addition to the intrinsic difficulties of coding medical texts mentioned above, clinical coding in Spanish requires dealing with the lack of textual resources crucial for training accurate automated systems. Hence, the lack of extensive training data in Spanish restricts the application of data-hungry DL methods, which have shown promising results in English clinical coding tasks. In the specific case of automatic coding of clinical texts in Spanish, there has been limited research.

The main objective of our work in [13] for the Text2RWD project was to develop clinical coding models for medical documents in Spanish, adapting several Transformer models to the particularities of the Spanish healthcare environment. To this end, the models were pre-trained with the Galén corpus [2]. The resulting models were refined on three clinical coding tasks from evaluation campaigns—CodiEsp-D, CodiEsp-P, and Cantemist-Coding [14, 15]—using two public Spanish annotated clinical corpora [15, 6, 16, 14]. In this work, three Transformer-based models that support the Spanish language were explored: multilingual BERT (mBERT), BETO and XLM-RoBERTa. In order to adapt the transformers to the particularities presented by clinical coding tasks with small datasets coming from the real world (clinical practice data stored in EHRs), a multi-label sentence classification approach was developed in this study and served as a DA procedure. Following the proposed strategy, the trained transformers achieved a new state of the art (SOTA) in each of the three clinical coding tasks explored in this work.

Table 1 shows the predictive results obtained in [13] for the three Transformer models analyzed: on the one hand, models without adaptation to the clinical-oncological domain (mBERT, BETO and XLM-R) and, on the other hand, models with adaptation to the clinical-oncological domain through pre-training with the Galén corpus (mBERT-Galén, BETO-Galén and XLM-R-Galén). The table also shows the results obtained by means of "expert committee" or "ensemble" strategies for the different models. As can be seen in the table, the best clinical coding results in the three tasks evaluated are obtained with the models adapted to the domain by pre-training with the Galén corpus. As expected, the model ensembles manage to improve the results obtained by the models independently.

2.4. Explainable clinical coding

Few studies have delved into the explainability of ML and DL models for clinical coding. In our work for the Text2RWD project [17], several approaches based on Transformer models, adapted to the clinical-oncology domain and multilingual, were developed and evaluated in order to address explainable clinical coding. These models have the ability to assign standardized disease and procedure codes to clinical documents, while providing detailed information about the specific text segments underlying the choice of each assigned code. To achieve this, the performance of two different multilingual Transformer models, namely XLM-RoBERTa and mBERT, as well as a Transformer-based model designed for the Spanish language, called BETO, were examined. These pretrained models were adapted to a specific clinical domain by continuous training with Galén corpus, and then refined and evaluated in subsequent clinical explanatory coding tasks: CodiEsp-X [14] and Catemist-Norm [15].

In addition, a comparison was made between two different training strategies for dealing with explainable clinical coding: a hierarchical approach versus a multitask approach. In the first approach, a Transformer is initially trained on a named medical entity recognition (MER) task to identify clinical entities, i.e. text fragments containing information relevant to medical diagnoses or procedures. The results of this MER Transformer are then used to train a second Transformer that deals with the medical entity normalization (MEN) task, assigning ICD-10 labels to the clinical entities previously identified by the first Transformer model. In the multi-tasking approach, the MER and MEN transformers are trained simultaneously.

The hierarchical approach (MER->MEN) for explainable clinical coding demonstrated better performance compared to the multi-task approach. Furthermore, the performance of domain-adapted transformers was found to outperform their non-adapted counterparts in all scenarios evaluated in this study (see Table 2). The various multilingual Transformer models and training approaches proposed in [17] were evaluated on public datasets obtained from shared tasks related to explanatory clinical coding, specifically CodiEsp-X [14] and Cantemist-Norm [15]. The results obtained in both tasks exceeded the SOTA for these tasks at the time of publication.

3. Conclusions and Future Work

This article presents the latest advances made by the ICB-UGCIO research group in the Text2RWD project. On the one hand, the progress made in the development of AI and DL algorithms—and, more specifically, in the design and training of models based on the RNN and the Transformer architecture—has demonstrated their usefulness and effectiveness in the performance of classification and named-entity recognition tasks for text classification, de-identification and explainable clinical coding in Spanish. Domain adaptation strategies—in which language models are trained using multilingual general purpose corpora and retrained using the Galén corpus, specific

Table 1
Performance (MAP metric) of the analyzed Transformer models for three clinical coding tasks (CodiEsp-Diagnoses, CodiEsp-Procedures and Cantemist-Coding), with and without adaptation to the clinical domain. The first six rows show the results of the independent models. The next six rows show the results obtained with ensembles of the previous models. The last row shows the SOTA at that time. The best results are shown in bold; the second best results are shown in underlined. (Extracted from [13]).

Ensemble	CodiEsp-D	CodiEsp-P	Cantemist-Coding
mBERT	.633	.508	.861
mBERT-Galén	.64	.521	.876
BETO	.625	.496	.853
BETO-Galén	.648	<u>.537</u>	<u>.88</u>
XLM-R	.629	.501	.862
XLM-R-Galén	.645	.526	.875
mBERT + mBERT-Galén	.65	.528	.874
BETO + BETO-Galén	.649	.534	.874
XLM-R + XLM-R-Galén	<u>.651</u>	.524	.874
mBERT + BETO + XLM-R	.647	.52	.87
mBERT-Galén + BETO-Galén + XLM-R-Galén	**.662**	**.544**	**.884**
Prior SOTA	.593	.493	.847

Table 2
Performance of the Transformer models analyzed for three explanatory clinical coding tasks (CodiEsp-X and Cantemist-Norm), with and without adaptation to the clinical domain, following a hierarchical training strategy. The first six rows show the results of the independent models. The next six rows show the results obtained with assemblies of the previous models. The last row shows the SOTA at that time. The best results are shown in bold; the second best results are shown in underline. (From [17]).

Model	Cantemist-Norm			CodiEsp-X		
	P	R	F1	P	R	F1
BETO	.836	.828	.832	.708	.542	.614
BETO-Galén	.838	.834	.836	.695	.557	.619
mBERT	.840	.835	.838	.711	.544	.616
mBERT-Galén	.843	.840	.842	.707	.556	.622
XLM-R	.835	.833	.834	.693	.546	.610
XLM-R-Galén	.843	.838	.840	.696	.560	.620
BETO + BETO-Galén	.845	.835	.840	.712	.556	.624
mBERT + mBERT-Galén	.846	.841	.844	.722	.552	.626
XLM-R + XLM-R-Galén	.845	.841	.843	.712	.559	.626
BETO + mBERT + XLM-R	.845	.839	.842	**.724**	.552	.626
BETO-Galén + mBERT-Galén + XLM-R-Galén	**.852**	**.847**	**.849**	.718	**.566**	**.633**
Prior SOTA	.824	.826	.825	.687	.562	.611

to the clinical-oncological domain—and the use of TL approaches, have allowed us to obtain SOTA results in several competitive tasks in the field of clinical natural language processing in Spanish. Our latest unpublished results, which show recent advances in normalization of clinical entities on standardized ontologies like Snomed-CT or UMLS, are promising and augur the continuity of the progress made within the Text2RWD project in a line of research that is still new and which will remain active in the coming years. Finally, in the coming years the Text2RWD project will also advance in new predictive tasks on RWD based on AI and NLP of EHRs—such as TNM staging in cancer—and their transfer to domains other than oncology in which the algorithms have been initially designed.

Acknowledgments

The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN) under project PID2020-116898RB-I00, from the Universidad de Málaga and Junta de Andalucía through grant UMA20-FEDERJA-045, from the Malaga-Pfizer consortium for AI research in Cancer - MAPIC, and from the Instituto de Investigación Biomédica de Málaga – IBIMA (all including FEDER funds)

References

[1] N. Ribelles, J. M. Jerez, D. Urda, J. L. Subirats, A. Márquez, C. Quero, L. A. Franco, Galén: Sistema de Información para la gestión y coordinación de procesos en un servicio de Oncología, RevistaeSalud 6 (2010).

[2] N. Ribelles, J. M. Jerez, P. Rodriguez-Brazzarola, B. Jimenez, T. Diaz-Redondo, H. Mesa, A. Marquez, A. Sanchez-Muñoz, B. Pajares, F. Carabantes, M. J. Bermejo, E. Villar, M. E. Dominguez-Recio, E. Saez, L. Galvez, A. Godoy, L. Franco, S. Ruiz-Medina, I. Lopez, E. Alba, Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients, European Journal of Cancer 144 (2021) 224–231. doi:10.1016/j.ejca.2020.11.030.

[3] D. Urda, N. Ribelles, J. L. Subirats, L. Franco, E. Alba, J. M. Jerez, Addressing critical issues in the development of an oncology information system, Int. J. Med. Inform. 82 (2013) 398–407.

[4] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of CLEF 2022, Bologna, Italy, 2022, pp. 179–203.

[5] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: Working Notes of CLEF 2023, Thessaloniki, Greece, 2023, pp. 1–18.

[6] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, ICB-UMA at CANTEMIST 2020: Automatic ICD-O Coding in Spanish with BERT, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020, pp. 468–476.

[7] G. López-García, J. M. Jerez, L. Franco, F. J. Veredas, Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data, PLOS ONE 15 (2020) 1–24. doi:10.1371/journal.pone.0230536.

[8] G. López-García, F. J. Moreno-Barea, H. Mesa, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Named Entity Recognition for De-identifying Real-World Health Records in Spanish, in: Computational Science – ICCS 2023, Springer Nature Switzerland, Cham, 2023, pp. 228–242.

[9] G. López-García, J. M. Jerez, L. Franco, F. J. Veredas, A Transfer-Learning Approach to Feature Extraction from Cancer Transcriptomes with Deep Autoencoders, in: Advances in Computational Intelligence, Springer International Publishing, Cham, 2019, pp. 912–924.

[10] D. Urda, F. J. Veredas, I. Turias, L. Franco, Addition of pathway-based information to improve predictions in transcriptomics, in: Bioinformatics and Biomedical Engineering: 7th International Work-Conference, IWBBIO 2019, Granada, Spain, May 8-10, 2019, Proceedings, Part II 7, Springer, 2019, pp. 200–208.

[11] G. López-García, J. M. Jerez, F. J. Veredas, ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT, in: Working Notes of CLEF 2020, CEUR Workshop Proceedings, 2020.

[12] F. J. Moreno-Barea, H. Mesa, N. Ribelles, E. Alba, J. M. Jerez, Clinical text classification in cancer Real-World data in spanish, in: Bioinformatics and Biomedical Engineering: 10th International Work-Conference, IWBBIO 2023, Meloneras, Gran Canaria, Spain, July 12–14, 2023, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 482–496.

[13] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Transformers for Clinical Coding in Spanish, IEEE Access 9 (2021) 72387–72397. doi:10.1109/ACCESS.2021.3080085.

[14] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at Codiesp track of CLEF ehealth 2020, in: Working Notes of CLEF 2020, Thessaloniki, Greece, 2020.

[15] A. Miranda-Escalada, E. Farré-Maduell, M. Krallinger, 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: Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Málaga, Spain, 2020, pp. 303–323.

[16] A. García-Pablos, N. Perez, M. Cuadros, Vicomtech at CANTEMIST, in: Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Málaga, Spain, 2020, pp. 489–498.

[17] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Explainable clinical coding with in-domain adapted transformers, Journal of Biomedical Informatics 139 (2023) 104323. doi:10.1016/j.jbi.2023.104323.