Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020

Antonio Miranda-Escalada[1], Aitor Gonzalez-Agirre[1], Jordi Armengol-Estapé[1], and Martin Krallinger[1]

Barcelona Supercomputing Center, Spain
antonio.miranda@bsc.es,aitor.gonzalez@bsc.es,jordi.armengol@bsc.es,
martin.krallinger@gmail.com

Abstract. Clinical coding requires the analysis and transformation of medical narratives into a structured or coded format using internationally recognized classification systems like ICD-10. These codes represent medical diagnoses and procedures. Clinical coding is critical for standardizing medical records, particularly for health information management systems used to carry out biomedical/ epidemiological research studies, monitor health trends or facilitate medical billing and reimbursement. The growing amount of clinical records has prompted the search for tools that assist manual coding. Inspired by the CCMC challenge and various eHealth CLEF shared tasks, we organized the CodiEsp track. Codiesp (eHealth CLEF 2020- Multilingual Information Extraction Shared Task) represents the first effort to promote the development and evaluation of automatic clinical coding systems for medical documents in Spanish. In this context, we have published a set of resources including (i) a manually coded Gold Standard corpus with inter-coder agreement and supporting textual evidence statements, (ii) an additional large collection of medical literature indexed with ICD-10 clinical codes and (iii) a machine translated corpus to enable multilingual approaches and testing of previous strategies developed for data in English. We have received a total of 168 runs submitted by 22 teams from 11 countries for at least one of our three sub-tracks: CodiEsp-D (Diagnosis Coding), CodiEsp-P (Procedure Coding) and CodiEsp-X (Explainable AI). Despite the considerable complexity of this task, which can be viewed as a hierarchical multi-label classification problem using ICD-10 codes as labels and documents as input, participants obtained very promising results, specially for codes that were well covered by the training data. Participants examined a variety of strategies, specifically deep learning approaches, pre-trained language models and word embeddings (BERT, BETO, FastText, etc.), as well as NER, string lookup and knowledge graph approaches. CodiEsp Corpus: https://zenodo.org/record/3837305

1 Introduction

Public health emergency situations, such as the COVID-19 global health crisis, further highlight the need of efficient search, retrieval, analysis, integration as well as exploitation strategies for a diversity of medical content types. This is particularly true for the medical literature, where clinical case reports characterizing in detail the symptoms and signs experienced by individual patients, together with the diagnosis, treatment and follow-up information, constitute a valuable evidence source for the possible pathogenesis of a disease and suitable therapeutic approaches [2]. Direct extraction of relevant information from electronic health records (EHRs) written by healthcare professionals represents a highly challenging problem due to (a) the rapid data accumulation and large data volumes (size, growth, performance, scalability problem), (b) the diversity of types, structures, formats and even character encodings in which clinical records are being produced (document standardization/harmonization problem), (c) the complex and rich medical domain specific vocabulary/terminologies and language characteristics being used (specialized domain problem) and (d) the diversity of languages and language variants in which clinical records are being written worldwide (multilingual content challenge).

Structured clinical information, in the form of coded clinical data relying on controlled indexing vocabularies such as ICD-10 [1]is a key resource for statistical analysis techniques applied to patient data [43]. The results of clinical coding activities are being used, for instance, as aggregated data to analyze retrospective and prospective aspects of information contained in electronic health records (EHRs). Clinical coding is a complex and time-consuming process, carried out by trained experts. This task requires the assignment of codes from a clinical classification (typically the 10th revision of the International Statistical Classification of Diseases and Related Health Problems or ICD-10) that essentially represents diagnoses and procedures associated to electronic health records.

The use of automatic systems to assist coding experts is becoming increasingly relevant to keep up with the pace of newly generated clinical texts. Automated clinical coding systems represent also a mechanism to improve coverage and consistency during the transformation process of EHRs into their corresponding structured representations.

Clinical natural language processing and AI-based document indexing strategies can result in resources useful for automatic clinical coding, directly exploiting the unstructured content of EHRs. Such tools play an increasing role to generate results that complement health informatics approaches focusing on translational medicine challenges, by providing relevant diagnostic information extracted from clinical narratives. This implies that text mining generated clinical coding results can provide a rich clinical context for patient health information necessary for other downstream data analysis processes like bioinformatics and OMICS data exploration. Figure **??** shows a general view of a canonical text mining flowchart and underlying tasks.

[1] https://www.who.int/classifications/icd/icdonlineversions/en/

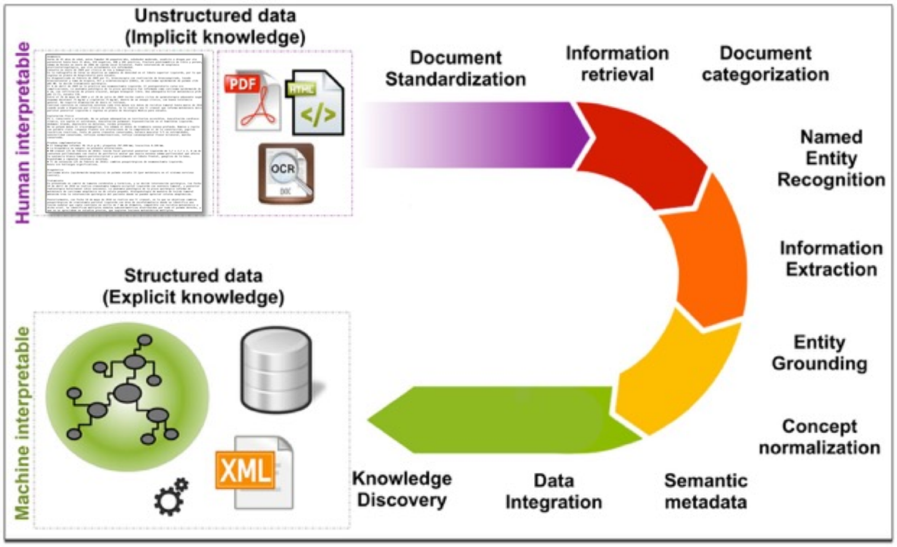


Fig. 1. Overview of canonical clinical text mining flowchart

Currently, most research on clinical NLP applies to English texts; however, there is a considerable amount of biomedical documents generated in non-English languages. The importance of clinical coding in languages other than English has driven challenges to promote automatic clinical coding systems. There have been community efforts to develop such clinical coding systems in English. In addition, in recent years, shared tasks for clinical coding have been proposed for English [37] as well as content in non-English languages such as French [15], German [11] or Japanese [4]. However, more limited research has been done in Spanish, despite the large volume of clinical content generated in this language, not only in Europe but worldwide. The success of these to competitions precedes CodiEsp, the first shared task on clinical case report coding in Spanish. The volume and growth rate of clinical texts written in Spanish worldwide justifies the need to promote not only the development of new text mining resources for clinical and medical narratives in Spanish, but also to carry out shared tasks and evaluation efforts to assure that the quality and used methods are competitive enough to be of practical value [1, 30, 24, 23]. Therefore, and also due to the interest in the health sector by the language technology industry, one of the flagship projects of the Spanish National Plan for the Advancement of Language Technology (Plan TL) is related to the clinical and biomedical field [48].

In the following sections, we will summarize the CodiEsp shared task setting, evaluation metrics, corpus preparation/annotation process, as well as the results produced by participating teams and a short summary of the used methodologies.

2 Task description

The CodiEsp [2] track proposes participants the challenge of building an automatic clinical coding system for Spanish documents. Participant systems have to automatically assign ICD-10 codes (CIE-10 in Spanish) to clinical case documents. Evaluation is done by comparing automatically generated results against manually manually generated ICD-10 codifications.

2.1 Subtasks

CodiEsp is structured into three different subtasks, two of them directly related to the two main branches of ICD-10 terminology. Moreover, to improve systems' acceptance, usefulness and practical integration into clinical coding support applications, results must be understandable, traceable to human-interpretable evidence sources and transparent. To that extent, in addition to two traditional coding subtasks, the CodiEsp shared task proposes a novel subtask on Explainable/Interpretable AI. Systems that participated in this subtask had to recognize the correct clinical codes and return the corresponding evidence text supporting the code assignment. The CodiEsp track comprised the following three subtracks:

- *CodiEsp Diagnosis Coding sub-task (CodiEsp-D)*: required automatic ICD-10-CM [CIE10 Diagnóstico] code assignment. This sub-track evaluated systems that predict ICD-10-CM codes (in the Spanish translation, CIE10-Diagnóstico codes). A list of valid codes for this sub-task with their English and Spanish description was provided by the task organizers [3].
- *CodiEsp Procedure Coding main sub-task (CodiEsp-P)*: required automatic ICD-10-PCS [CIE10 Procedimiento] code assignment. This sub-track evaluated systems that predict ICD-10-PCS codes (in the Spanish translation, CIE10-Procedimiento codes). A list of valid codes for this sub-task with their English and Spanish description was provided by the task organizers.
- *CodiEsp Explainable AI exploratory sub-task (CodiEsp-X)*. Participating systems were asked to return in addition to clinical code assignments supporting evidence texts extracted from documents. Both ICD-10-CM and ICD-10-PCS codes were used for this subtask. Evaluation was done against a collection of manually labeled evidence texts.

2.2 Shared task setting and schedule

The CodiEsp track was organized in the form of three basic participation periods:

1. Training phase. During the initial participation period, a random subset of the entire corpus was published corresponding to the training data collection.

[2] https://temu.bsc.es/codiesp

[3] https://zenodo.org/record/3706838

Table 1. Example showing a comparison between automatic clinical ICD-10 code predictions and manual coding annotations for the CodiEsp-Diagnostic and CodiEsp-Explainability sub-tracks.

Evidence text annotation	Automatic prediction	Manual coding	Comparison
	Diagnostic		
El paciente refiere dolores osteoarticulares de localización variable	m25.50	m25.50	✓
	Explainability		
English: The patient complained of osteoarticular pain of variable location	m25.50 ”osteoarticulares”	m25.50 “dolores osteoarticulares”	x

It consisted of plain text documents and their corresponding annotations, i.e. ICD-10 code assignments and manually labeled evidence texts. During this period, teams started implementing their automatic clinical coding strategies by exploiting this dataset.

2. Development phase. Next, a second subset of the corpus was released (development data). This dataset served to fine tune and improve the initial predictive coding systems.
3. Test phase. Finally, the test set was released. This third subset of the corpus was distributed without providing manual annotations/code assignments. Participants had to return for all test set documents their corresponding ICD-10 codes. After the submission deadline, the shared task organizers evaluated team predictions against manual code assignments/annotations. A total of 5 runs were allowed for each subtrack per team, so that participants could explore different strategies and methodological approaches.

2.3 Evaluation metrics used for CodiEsp

In case of the CodiEsp-Diagnostic and CodiEsp-Procedure subtasks, automatic predictions returned by teams had to consist in ranked codes. The primary evaluation metric for these two subtasks was Mean Average Precision (MAP).

Mean Average Precision (MAP) is a widely used evaluation score for ranking problems:

$$AveP = \frac{\sum(P(k) * rel(k))}{\text{number of relevant documents}}$$

where, P(k) is the precision at the position k, and rel(k) is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

MAP has shown good discrimination and stability [29]. For completeness, error analysis, and comparison to previous efforts, other metrics were also computed: MAP@k (MAP taking into account just the first k results), f1-score, precision, and recall.

Table 2. List of metrics computed in each subtask.

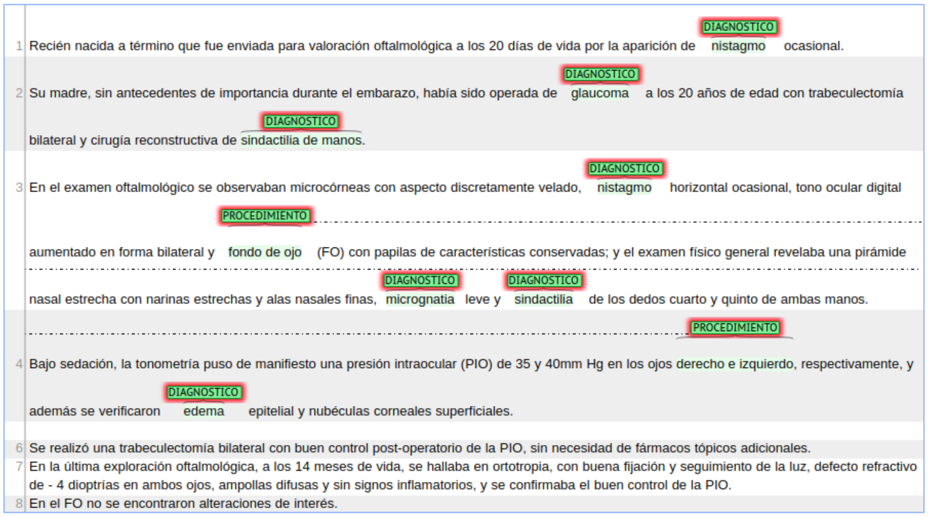
	CodiEsp-Diag.	CodiEsp-Proc.	CodiEsp-Exp.	Metric description
MAP	✓	✓	x	Mean Average Precision
MAP@30	✓	x	x	MAP considering the first 30 predictions per document
MAP@10	x	✓	x	MAP considering first 10 predictions per document
MAP train and dev codes	✓	✓	x	MAP considering only codes present in training and development sets
P,R,F1	✓	✓	✓	Precision, Recall and F1-score
P,R,F1 train and dev codes	✓	✓	✓	P, R, and F1 considering only codes present in training and development sets
P,R,F1 categories	✓	✓	x	P, R, and F1 considering only the first 3 digits of the codes (4 in procedure codes)

Participants of the CodiEsp-Explainability subtask were evaluated with micro balanced f1-score, precision, and recall since its scope is different and more complicated.

$$\begin{aligned}
\text{Precision (P)} &= \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \\
\text{Recall (R)} &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\
\text{F1 score (F1)} &= \frac{2 * (P * R)}{(P + R)}
\end{aligned}$$

A complete overview of all used evaluation metrics for the CodiEsp track is shown in Table 2. In addition, Table 1 shows an evaluation example.

CodiEsp baseline system. To provide context to the obtained task results, we implemented a baseline system using dictionary lookup and vocabulary transfer. This baseline system selects manually labeled text spans from the training and development collections, using these mentions afterwards as a gazetteer for lexical looking up in the test set documents. The lookup was strict. However, texts had been previously tokenized and normalized (transformed to lowercase, accents were removed, extra blank spaces or punctuation signs were ignored).



1 Recién nacida a término que fue enviada para valoración oftalmológica a los 20 días de vida por la aparición de nistagmo ocasional.

2 Su madre, sin antecedentes de importancia durante el embarazo, había sido operada de glaucoma a los 20 años de edad con trabeculectomía bilateral y cirugía reconstructiva de sindactilia de manos.

3 En el examen oftalmológico se observaban microcórneas con aspecto discretamente velado, nistagmo horizontal ocasional, tono ocular digital aumentado en forma bilateral y fondo de ojo (FO) con papilas de características conservadas; y el examen físico general revelaba una pirámide nasal estrecha con narinas estrechas y alas nasales finas, micrognatia leve y sindactilia de los dedos cuarto y quinto de ambas manos.

4 Bajo sedación, la tonometría puso de manifiesto una presión intraocular (PIO) de 35 y 40mm Hg en los ojos derecho e izquierdo, respectivamente, y además se verificaron edema epitelial y nubéculas corneales superficiales.

6 Se realizó una trabeculectomía bilateral con buen control post-operatorio de la PIO, sin necesidad de fármacos tópicos adicionales.

7 En la última exploración oftalmológica, a los 14 meses de vida, se hallaba en ortotropia, con buena fijación y seguimiento de la luz, defecto refractivo de - 4 dioptrías en ambos ojos, ampollas difusas y sin signos inflamatorios, y se confirmaba el buen control de la PIO.

8 En el FO no se encontraron alteraciones de interés.

Fig. 2. Annotated clinical case report visualized with Brat tool [44].

3 Corpora and Resources

3.1 CodiEsp corpus

The CodiEsp corpus[4] is a collection of 1,000 clinical case reports written in Spanish that cover a diversity of medical specialities. The training subset consisted in 500 documents, while the development and test set consisted in 250 documents each. All documents were exhaustively, manually annotated by professional clinical coders with codes from the Spanish version of ICD-10 (procedure and diagnostic). Additionally, human annotators had to label or mark up clinical-coding evidence text fragments. Figure 2 shows an example document with manually annotated codes, and Table 3 shows an example of an annotated sentence in Spanish and in English.

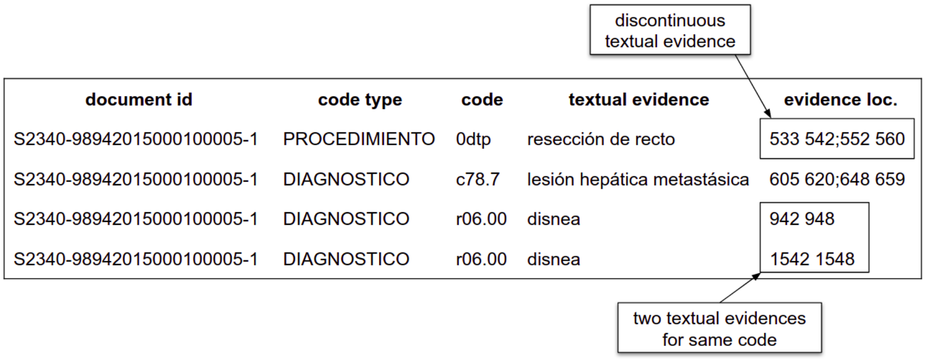
The manual annotation process followed official clinical coding guidelines published for Spain. CodiEsp documents were coded with the 2018 version of CIE-10 (the official Spanish version of ICD-10-Clinical Modification and ICD-10-Procedures) and inspired by the "Manual de Codificación CIE-10-ES Diagnósticos 2018" and the "Manual de Codificación CIE-10-ES Procedimientos 2018" provided by the Spanish Ministry of Health. To cover aspects and particularities relevant to the sub-tasks and documents used for CodiEsp, together with the corpus itself, a document describing the annotation guidelines was published [5]. Clinical codes (diagnostic and procedure) were linked to textual evidence fragments that support their assignment (see Figure 3).

[4] `https://doi.org/10.5281/zenodo.3625746`

[5] https://zenodo.org/record/3730567

Table 3. Example of an annotated sentence.

	Code	Textual evidence
Spanish: El paciente presentó un cuadro brusco de disnea, vómitos y pérdida de conocimiento. *English automatic translation:* The patient developed sudden dyspnea, vomiting, loss of consciousness	r55 r06.00 r11.10	disnea vómitos pérdida de conocimiento



discontinuous textual evidence

document id	code type	code	textual evidence	evidence loc.
S2340-98942015000100005-1	PROCEDIMIENTO	0dtp	resección de recto	533 542;552 560
S2340-98942015000100005-1	DIAGNOSTICO	c78.7	lesión hepática metastásica	605 620;648 659
S2340-98942015000100005-1	DIAGNOSTICO	r06.00	disnea	942 948
S2340-98942015000100005-1	DIAGNOSTICO	r06.00	disnea	1542 1548

two textual evidences for same code

Fig. 3. Example of tab-separated file.

To guarantee annotation quality, we performed an iterative process of guideline refinement and consistency analysis through comparison between independent annotations provided by multiple clinical coders. Several initial annotation rounds were necessary until an acceptable level of manual annotation quality was obtained. For the annotation of diagnostic codes, the final pairwise percentage agreement obtained was 88.6%, 88.9% for procedure codes and 80.5% for the annotation of textual evidence.

Corpus format. Gold Standard CodiEsp corpus is distributed in the CodiEsp format: documents are provided in plain text format, and annotations are released in a tab-separated file. Each line of the file corresponds to a code assignment. This format is coherent with the data format used in the 2019 CLEF clinical coding shared task [11]. See Figure 3 for an example of the tab-separated file with annotation information. In it, there are examples of discontinuous textual evidence and codes with more than one textual evidence.

Corpus statistics. In total, the entire CodiEsp corpus contains 18435 annotations, with the DIAGNOSTIC class more common than the PROCEDURE class: 77.8% of the annotations correspond to diagnostics. The 18435 annotations contain 3427 unique ICD-10 codes. Again, there are more diagnostic than

Table 4. Summary statistics of CodiEsp corpus.

	Doc.	**Annotations**			**Unique codes**			**Sentences**	**Tokens**
		diagnostic	procedure	total	diagnostic	procedure	total		
Train	500	7209	1972	9181	1767	563	2330	8105	204815
Dev.	250	3431	1046	4477	1158	375	1533	4381	102719
Test	250	3665	1112	4777	1143	371	1514	4198	103533
Total	1000	14305	4130	18435	2557	870	3427	16684	411067

procedure codes: 2557 and 870, respectively. We hypothesise that it is more complicated for an automatic system to predict procedures, since the corpus contains fewer examples. However, since there are also less unique procedure codes, this difficulty may be partially addressed. Table 4 contains a summary of the corpus statistics.

3.2 Additional resources

We have generated a collection of additional resources to overcome the size limitation of our Gold Standard CodiEsp corpus. These resources included:

CodiEsp MT corpus [6]. The CodiEsp shared task attracted participants from many non-Spanish speaking countries. In addition, there are already many clinical coding systems for data in English. To ease the comparison with such systems, provide support to participants working previously in English and to explore the use of machine-translated corpora, we generated a machine translated version of the CodiEsp corpus (CodiEsp MT corpus). The used machine translation system was adapted to the language characteristics of the medical domain [42].

CodiEsp-abstracts [7]. To increase the size and number of possible training instances, we prepared a dump of medical literature abstracts from the Lilacs [26] and IBECS [18] bibliographic resources. Those were indexed manually with either DeCS or MeSH terms. Using a mapping chain [DeCS → MeSH → UMLS → ICD-10], we generated a collection of medical literature abstracts with associated ICD-10 codes. The resulting collection contains 176,294 Spanish medical abstracts indexed with ICD-10 codes.

A mapping chain was generated. DeCs is a terminological resource created to index journal articles, technical reports, and other health-related documents. It is based on MeSH (Medical Subject Heading), developed by the U.S. National Library of Medicine [27]. Additionally, we used the UMLS Metathesaurus tool [47] to map MeSH terms to ICD-10 codes (see Figure 4). This enabled us to build a mapping from DeCS terms to ICD-10 codes with this mapping chain [DeCS → MeSH → UMLS → ICD-10].

[6] `https://doi.org/10.5281/zenodo.3625746`

[7] `https://doi.org/10.5281/zenodo.3606625`

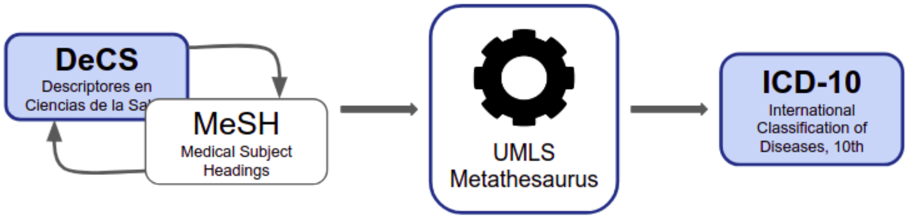

DeCS
Descriptores en Ciencias de la Sa
MeSH
Medical Subject Headings
UMLS Metathesaurus
ICD-10
International Classification of Diseases, 10th

Fig. 4. DeCs to ICD-10 mapping.

PubMed machine-translation [8]. A large collection of PubMed abstracts was automatically translated into Spanish using the same translation engine employed to translate the CodiEsp corpus [42]. PubMed abstracts are manually indexed with MeSH and easily mapped to ICD-10 terms using the same mapping chain employed for CodiEsp-abstracts. This resource was also provided to participants.

CodiEsp Silver Standard [9]. The CodiEsp test set documents were released together with an additional collection of 2,751 clinical case documents (called the background set). Participants were asked to provide code predictions for the entire collection of 3,001 documents (background set plus test set). Such a setting tried to examine if participating systems were able to scale to larger data collections. Code predictions for the background set were released as a CodiEsp Silver Standard corpus, similar to the CALBC initiative [40].

4 Results

4.1 Participants description

We received submissions from a total of 22 teams. In the CodiEsp-Diagnostic subtask, there were 22 participants (78 runs). For the CodiEsp-Procedure track, we received 64 runs from 17 teams. The exploratory CodiEsp-Explainability subtask had 8 participants, which returned a total of 25 runs. In total, 167 novel clinical coding systems were generated in the context of CodiEsp. These numbers are shown in Table 6

A detailed description of the participant teams is included in Table 5.

4.2 Systems results

Table 7 shows the results of the best run obtained by each team. The top scoring results for each subtask were:

[8] `https://doi.org/10.5281/zenodo.3826553`
[9] `https://doi.org/10.5281/zenodo.3859869`

Table 5. CodiEsp team overview. A/I stands for academic or industry institution. In the Tasks column, D stands for CodiEsp-Diagnostic, P for CodiEsp-Procedure and E for CodiEsp-Explainability.

Team Name	**Affiliation**	**A/I**	**Tasks**	**Ref.**	**Tool URL**
TeamX	Fuji Xerox Co., Ltd., Japan	I	D,P	[46]	-
SWAP	University of Bari, Italy	A	D,P	[38]	[45]
LIIR	KU Leuven, Belgium	A	D,P	[32]	-
FLE	Fujitsu Laboratories of Europe, Spain	I	D,P,E	[14]	-
IAM	ERIAS, France	A	D,P,E	[7]	[17]
BCSG	University of Applied Sciences and Arts Dortmund, Germany	A	D	[41]	-
SINAI	University of Jaén, Spain	A	D,P,E	[36]	-
DCIC - UNS	Universidad Nacional del Sur, Argentina	A	D,P	-	-
IMS	University of Padua, Italy	A	D,P	[34]	[21]
SSN-NLP	SSN College of engineering, India	A	D,P	[25]	-
MEDIA	University of the Basque Country, Spain	A	D,P	[20]	-
Hulat-PDPQ	University Carlos III, Spain	A	D	[39]	[16]
NLP-UNED	National Distance Education University, Spain	A	D,P	-	-
UDC-UA	University of Aveiro and University of A Coruña, Portugal	A	D,P,E	-	-
CodeICD@IITH	Indian institute of technology Hyderabad, India	A	D	-	-
The Mental Strokers	Fraunhofer Portugal AICOS, Portugal	-	D,P,E	[9]	-
ICB-UMA	University of Málaga, Spain	A	D	[28]	[19]
ExeterChiefs	University of Exeter, UK	A	D,P	[35]	[13]
LSI-UNED	UNED, Spain	A	D,P	[3]	-
nlp4life	Data4Life, Germany	A	D	[12]	[33]
Anuj	Sapient, US	I	D,P,E	-	-
IXA-AAA	University of the Basque Country, Spain	A	D,P,E	[5]	-

Table 6. Summary of participation results of CodiEsp.

	CodiEsp-Diagnostic	**CodiEsp-Procedure**	**CodiEsp-Explainability**	**Total**
Participant teams	22	17	8	22
Submitted runs	78	64	25	167

- *CodiEsp-Diagnostic.* IXA-AAA, reached a MAP of 0.593. They obtained a high recall and a moderate precision. The highest f1-score was achieved by the IAM team, with 0.817 precision and 0.592 recall.

Table 7. Precision, recall, f1-score, and MAP of best run for CodiEsp-Diagnostic, CodiEsp-Procedure and CodiEsp-Explainability. Bolded, the best result, underlined the second-best.

	CodiEsp-Diag.				**CodiEsp-Proc.**				**CodiEsp-Expl.**		
Team Name	**P**	**R**	**F1**	**MAP**	**P**	**R**	**F1**	**MAP**	**P**	**R**	**F1**
TeamX	.123	.858	.192	.299	.042	**.825**	.077	.19	-	-	-
SWAP	.295	.442	.308	.202	.186	.513	.25	.221	-	-	-
LIIR	.124	.055	.076	.044	.051	.034	.041	.02	-	-	-
FLE	.74	.633	.679	.519	.643	.448	.514	.443	.687	**.562**	**.611**
IAM[10]	.817	.592	**.687**	.521	.691	.42	**.522**	**.493**	**.75**	.524	**.611**
BCSG	.457	.287	.337	.259	-	-	-	-	-	-	-
SINAI	.45	.544	.488	.314	.37	.476	.416	.293	.36	.447	.399
DCIC - UNS	.482	.261	.187	.097	.471	.074	.082	.105	-	-	-
IMS	.373	.709	.474	.449	.31	.749	.376	.391	-	-	-
SSN-NLP	.025	.049	.033	.007	.016	.075	.027	.028	-	-	-
MEDIA	.735	.63	.629	.488	.601	.52	.502	.442	-	-	-
Hulat-PDPQ	**.866**	.066	.123	.115	-	-	-	-	-	-	-
NLP-UNED	.542	.089	.153	.1	.34	.018	.035	.168	-	-	-
UDC-UA	.727	.605	.546	.368	.82	.34	.357	.33	.678	.492	.463
CodeICD@IITH	.462	.281	.35	.192	-	-	-	-	-	-	-
The Mental Strokers	.759	.638	.591	.517	.537	.527	.488	.445	.534	.478	.505
ICB-UMA	.004	**.897**	.009	.482	-	-	-	-	-	-	-
ExeterChiefs	.117	.201	.144	.082	.106	.325	.145	.125	-	-	-
LSI-UNED	.253	.688	.37	.517	.066	.569	.119	.398	-	-	-
nlp4life	.014	.038	.02	.004	-	-	-	-	-	-	-
Anuj	.741	.621	.676	.505	**.833**	.396	.478	.413	.572	.456	.507
IXA-AAA	.004	.858	.009	**.593**	.004	**.825**	.008	.425	.288	.318	.283

- *CodiEsp-Procedure.* The IAM team achieved the highest MAP, 0.493, and f1-score, 0.522. The precision of this team was 0.691, and the recall was 0.42.
- *CodiEsp-Explainability.* The top-performing team was FLE, with the best f1-score (0.611). It obtained a precision of 0.687, and its recall 0.562. In an unofficial run, the IAM team obtained the same f1-score. Since this subtask required identifying not only the right codes, but also the correct textual evidence, MAP metric was not computed.

Codes present in training and development. The division of the CodiEsp corpus into training, development, and test set was performed using randomly generated non-overlapping samples. Since the ICD-10 terminology has more than 170.000 distinct codes, some codes present in test documents were not covered previously by training or development set annotations. When evaluating systems using only the subset of codes present in training and development sets, all evaluation scores increase considerably. In the Figure 5, this effect is clearly observed. In red, we have the f1-score values computed taking into account all codes. In blue, we have the same metric taking into account just the codes present

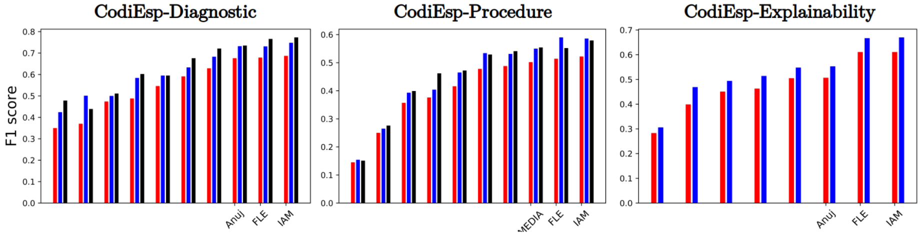


Fig. 5. F1 comparison. In red, main test set results. In blue, test set results considering only codes that were present in training and development sets. In black, test set results computed from code categories.

in training and development sets. Three teams in CodiEsp-Diagnostic subtask developed systems with an f1-score above 0.7. For a complete relation of metrics computed evaluating these codes, see Table 9, Table 10 and Table 11.

Code categories. Figure 5 also includes a third f1-score value in black. It corresponds to the same metric computed on the categories. ICD-10 terminology is tree-shaped (for Diagnostics) and axial (for Procedures). This characteristic means that the first digits of the code give different information than the last digits. As digits are located more to the right of the code, their information is more granular. Therefore, f1-score, precision, and recall were computed taking into account only the first three digits for CodiEsp-Diagnostic and the first four digits for CodiEsp-Procedure.

Systems that correctly tag the category but fail on the more granular information could be a starting point that requires fine-tuning. And this is the case of most participant systems of CodiEsp (Figure 5, Table 9, Table 10 and Table 11). Indeed, when observing these metrics, we can see that the best prediction run of the IAM team reaches 0.773 f1-score.

For a complete list of all metrics for all runs, check the Table 9, Table 10 and Table 11 at Appendix.

4.3 Error analysis

In this error analysis, the focus was placed on codes that, despite being present more than four times in the training or development sets, were predicted correctly by less than 20% of the runs. Such codes could be considered as "difficult".

Difficult codes have discontinuous textual evidences. Codes with discontinuous text evidence were, in general, more difficult to detect correctly. As seen in Figure 6, in the training, development, and test sets the percentage of codes with discontinuous textual evidence is below 15%. However, in the subset of difficult codes, it approaches 30%.

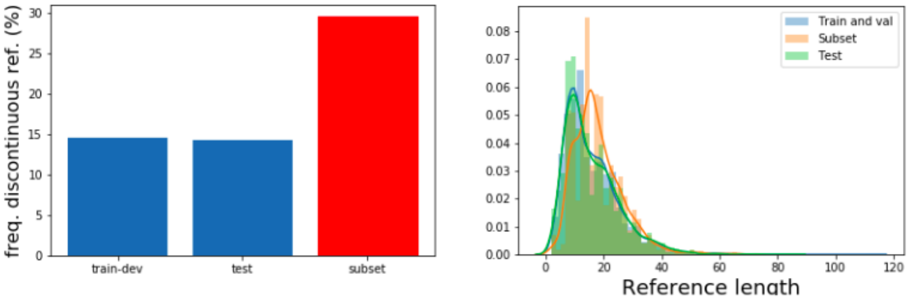


Fig. 6. Frequency of discontinuous references and reference length comparison in the training and development set, the test set and the subset of interest.

Besides, it is clear from Table 7 that results for CodiEsp-Procedure are worse than those for CodiEsp-Diagnostic. In the former, 38.7% of the textual evidence of the test sets are discontinuous. In the latter, 14.3%.

Additionally, codes that are well predicted by most teams include continuous pieces of evidence. In fact, among the codes successfully predicted in more than half of the runs, the proportion of discontinuous texts of evidence is 4.3%, while this proportion increases to 14.3% in the whole test set.

Difficult codes have longer textual evidences. Not only codes with discontinuous references are difficult to detect. Also, codes whose textual reference is longer are more challenging. In Figure 6, we observe how the distribution of textual reference lengths is almost identical for the codes in the training, development, and test sets (blue and green). However, for our subset of codes of interest, the average length increases.

Less specific codes are predicted with higher accuracy. In diagnostics, codes ending with a 9 tend to be less specific than the others. For example, code *M25.561* represents ”pain in the right knee”, while *M25.569* is used for ”pain in an unspecified knee”. The same happens with procedure codes ending with Z. We have evaluated the codes predicted by more than 50% of participant runs. Those codes could be seen as “easier” codes since most systems assign them. And there are 34.2% of 9-ending codes in the training and development sets, 36.8% of such codes in the test set, while there are 52.5% of such codes in the “easy” subset of codes.

There are more abbreviations in procedures than in diagnostics. Abbreviations are a common problem when processing medical narratives. Their presence is ubiquitous in this type of texts, and their meaning varies from one medical specialty to another. We have looked for Spanish medical abbreviations, collected in the Spanish Medical Abbreviation DataBase [22], in the textual pieces of evidence that justify the code assignment.

Abbreviations appear in the difficult subset of codes in a similar percentage as in the rest of the corpus. For instance, 18.9% of diagnostic code evidence have abbreviations, and that percentage in the difficult codes is 19.1%. However, abbreviations appear much less in a subset of "easy" codes (codes predicted by more than 50% of the runs), 13%. For procedures, the phenomenon is the same. Indeed, there are more abbreviations in the procedure than in the diagnostic textual evidences: 40% of procedure textual evidences contain abbreviations from the Spanish Medical Abbreviation DataBase, against 19.1% for diagnostic textual evidences. This might contribute to the difficulty of assigning procedure codes.

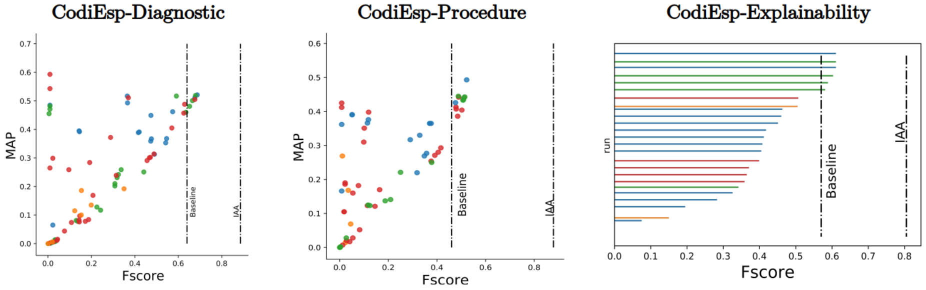
5 Participant Methodologies

Methodology distribution. With participants from diverse backgrounds, the range of methodologies employed is broad. For specific details of the participants' systems, we refer you to the particular articles. However, a simplified classification is presented in this paper. Participant systems were divided into those that employ language models (such as the popular BERT [10]), those that integrate other Machine Learning algorithms, and those that do not use Machine Learning. MAP and f1-score (taking into account codes present in training and development test) results are shown in Figure 7 colored by methodology.

In the three subtasks, there are successful and unsuccessful teams in the three methodological groups. For example, the highest MAP in CodiEsp-Diagnostic is obtained by a team employing machine learning. In contrast, the top MAP scores for CodiEsp-Procedure was obtained by non-machine learning systems. The second-best f1-score in both subtasks was obtained by a fine-tuning Multilingual BERT (a language model) approach. Finally, it is noteworthy that in case of the CodiEsp-Explainability track, more teams were using non-machine learning strategies.

Participants descriptions. There have been three main approaches to automatic clinical coding:

- *Classification.* This approach considers that there are a set of documents that must be categorized. Every ICD-10 has its own category. For example, team ICB-UMA [28] followed this classification schema.
- *Named Entity Recognition.* In this case, automatic systems must detect whether each clinical case word (or set of words) is a diagnostic, a procedure, or none of them. Examples of NER systems are FLE [14] or IAM [7].
- *Combination.* For example, IXA-AAA [5] team combined a classifier and a NER system.


CodiEsp-Diagnostic

CodiEsp-Procedure

CodiEsp-Explainability

Fig. 7. MAP and f1-score values of all prediction runs. They are colored by the methodology used. Green is employed for systems using deep learning language models. Red for other Machine Learning algorithms. Blue for not Machine Learning. And orange when the methodology is unknown. Also, IAA, computed as the pairwise agreement, is included to compare the data with the human agreement. Finally, baseline f1-score results are also included. .

Additionally, each of the two schemas (classification vs. NER) may be tackled using different technologies. In this overview, we have clustered the technologies in 3 classes:

- *Non-machine learning approaches.* For instance, IAM [7] employs a dictionary lookup to perform NER.
- *Machine learning approaches.* In this group, we find IXA-AAA [5], that use XGBoost to perform document classification.
- *Language models.* Within the teams using machine learning, a significant number of them employ language models. For example, the FLE team [14] fine-tuned BERT Multilingual and The Mental Strokers [9] BETO language model.

In the following paragraphs, the approaches followed by some of the teams are briefly described to illustrate the different methodologies. Descriptions of other participant teams are found in their system description papers.

- *IXA-AAA.* They combined a Machine Learning engine with a string similarity system. First, a binary XGBoost classifier was trained for each label. Since the XGBoost outputs a probability for each prediction, its output could be used directly in subtasks CodiEsp-Diagnostic and CodiEsp-Procedure. Texts were expanded to improve the XGBoost models, concatenating the medical entities extracted from the documents itself. Second, the string similarity system compares text fragments and ICD-10 code definitions using Levenshtein distance, Jaro Winkler algorithm, and Cosine Similarity on Multilingual BERT representations. ICD-10 standard definitions were expanded with non-standard terms, single-word descriptions, and even phrases frequently associated with the codes. The best system for CodiEsp-Diagnostic and CodiEsp-Procedure subtasks is a combination of XGBoost and Jaro

Winkler string similarity outputs. This combination obtained the highest MAP for the CodiEsp-Diagnostic subtask, 0.593 [5].

- *IAM.* IAM team has employed the same clinical coding system in past clinical coding shared tasks [8]. Their system is based on a dictionary with a tree data structure built from CodiEsp annotations and ICD-10 terminology. Entities are detected in new documents if they match any of the stored entries of the dictionary. The match is performed by exact matching, Levenshtein matching, and abbreviation matching. This last matching modality uses a dictionary of abbreviations. To further improve the precision of the system, they removed terms that lead to many false positives. Their system obtained the highest f1-score in CodiEsp-Diagnostic and in CodiEsp-Procedure, 0.687 and 0.522, and the largest MAP in CodiEsp-Procedure, 0.493. Finally, it also achieved the highest f1-score in CodiEsp-Explainability (in an unofficial run), 0.611 [7].
- *The Mental Strokers.* They re-trained the BETO language model [6] on the training set of CodiEsp until the perplexity stabilized. Next, they added a linear classification layer and fine-tuned the model to perform NER on the CodiEsp corpus. That had to be recognized corresponded to the codes present in the training and development sets. A model for diagnostics and a different one for procedures were created. Also, they tested and submitted a system based on a Conditional Random Field (CRF) to perform the same NER task but concluded that it was too conservative in entity detection. With the fine-tuned language model, they obtained 0.445 MAP in CodiEsp-Procedures (the second position) [9].
- *FLE.* The FLE team used a two-step approach to detect ICD-10 codes in documents. First, diagnostic and procedure entities were identified using a NER system based on a pre-trained multilingual BERT. Second, the recognized entities were matched to ICD-10 code definitions or examples using Levenshtein distance. Also, they used an in-house text augmentation algorithm to increase the size of the training dataset artificially. The text augmentation algorithm was trained with the CodiEsp corpus, and examples from PubMed and MIMIC database translated to Spanish. Additionally, they tested different post-processing methods to detect and remove negated and overlapping entities. Finally, to rank the codes according to confidence, they used entity frequency and position: they considered that more frequent entities and entities mentioned closer to the end of the document were more likely to be correct. Their system obtained the highest f1-score in CodiEsp-Explainability, 0.611. Remarkably, FLE is one of the three participant teams from a commercial organization (Fujitsu, Spain) [14].

Combined methodologies. Since teams had approached the challenge from different perspectives, we analyzed what would happen if we combined predictions from different approaches. This rationale was already followed by the IXA-AAA team when they combined an XGBoost classifier with a string matching system. For instance, we may combine predictions of the FLE (that fine-tuned

Table 8. Precision, Recall, and F1-score comparison of IAM, FLE, and their combined (union) system.

	Precision	**Recall**	**F1**
IAM	**0.817**	0.592	0.687
FLE	0.739	0.556	0.635
IAM∪FLE	0.716	**0.691**	**0.703**

mBERT for NER and Levenshtein distance to find ICD-10 code) and the IAM (that employed a tuned dictionary lookup) teams. The combination is an uncomplicated union. Any code predicted by any of the two systems is considered.

Compared with the manual gold standard, we observe in Table 8 how the f1-score increases, and the resulting prediction is more balanced than the previous two. Both systems lacked a high recall, and combining them makes recall to increase. Precision decreases, but not as much. Then, the result is a prediction higher than any of the individual participant predictions.

6 Discussion

The CodiEsp task (eHealthCLEF 2020 Task 1 on Multilingual Information Extraction), has attracted a considerable number of participants. There were only 3 CLEF subtasks with more registrations (ImageCLEF Task 3, medical; Check-That Task 1, Check-Worthiness on tweets; and LifeCLEF Task 2, BirdCLEF). Indeed, participant teams came from very diverse countries, despite the highly specialized and complex application domain (medical/clinical data) and the use of non-English textual data. In this sense, CodiEsp could be perceived as a more challenging task setting when compared to other CLEF competitions, such as ImageCLEF or LifeCLEF.

Novelty and Impact. To the best of our knowledge, the CodiEsp corpus is the first publicly available, manually annotated, clinical coding gold standard text corpus in Spanish. Despite the use of commercial tools for automatic clinical coding in hospitals and private companies, a setting for direct comparison and benchmarking using a common evaluation set, format and metrics was missing. CodiEsp has partially solved this problem. However, there is still a need for large data collections, including high quality manually coded anonymized EHRs from multiple hospitals.

Access to patient reports is a recurrent problem in clinical NLP due to privacy issues. Creative solutions, such as using clinical cases that have been carefully selected as a surrogate data of real clinical reports, might contribute to the development of clinical NLP infrastructures. In this sense, we have also employed document and individual sentence similarity (lexical/surface as well as semantic similarity) strategies comparing real EHRs (mainly discharge summaries and radiology reports) with clinical case reports to retrieve sentences that are basically equivalent for corpus construction and public release purposes. We call this

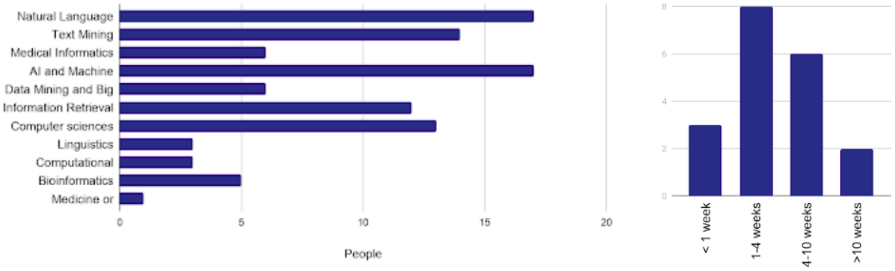


Fig. 8. Fields of knowledge of CodiEsp participants and the time invested in the shared task.

kind of creative corpus construction, circumventing data privacy issues as the **wandering corpus strategy**, where data with legal redistribution issues can at least be emulated by publicly available data resources that are highly similar.

CodiEsp is a clinical coding shared task in Spanish. Despite this apparent language constraint, there have been participants from 9 non-Spanish speaking countries. One of the reasons may be that 57.9% of participants reported that their system is multilingual and not Spanish-specific. CodiEsp not only attracted participants from different countries, but also different backgrounds (Figure 8). Even though Natural Language Processing, Artificial Intelligence, and Machine Learning were the most numerous background, there were also participants coming from fields such as bioinformatics, linguistics, and medicine or biomedicine. Finally, 57.9% of the participants reported experience in clinical coding systems (or similar tasks) before CodiEsp, and most of them said that their motivation was "to be able to compare their results with other strategies/methods/teams."

The participation of such diverse profiles has allowed the creation of heterogeneous resources, available to the community. Such resources are centralized in the CodiEsp webpage[11]. In this sense, 68.4% of the participant teams indicated that they would provide software or web service based on their system if there is specific technical or financial support, whereas 31.6% were not interested in assistance to advance their clinical coding system into a software product or startup.

CodiEsp has been a challenging shared task according to the participants. Most of them rated it as difficult or very difficult. However, 78.9% would be interested in participating in a second CodiEsp track. When asked about the time invested, the most common answer has been 1 to 4 weeks, followed by 4 to 10 weeks (Figure 8).

Possible improvements. One of the CodiEsp limitations has been the Gold Standard size. CodiEsp corpus contains 1,000 annotated clinical case reports, with 16,504 sentences and 396,988 tokens. In the shared task setting, 750 docu-

[11] `https://temu.bsc.es/codiesp`

ments were used for learning, and 250 were employed for results evaluation. As previously discussed, some codes present in the test set had not been employed in the training and development sets. Indeed, 1856 codes appear just once in the entire gold standard. Even though several systems achieved high metrics, a more significant Gold Standard would allow more examples for systems to learn from, and a more representative set of documents to evaluate them. An extension of the CodiEsp corpus would now require fewer resources, since the guidelines employed are already publicly available [31].

Additionally, as some participants have pointed out, the gold standard of the exploratory subtask CodiEsp-Explainability had a limited inter-annotator agreement. CodiEsp corpus was annotated independently by two clinical experts. On the subset of documents annotated by both experts, they achieved an IAA of 80.5% for annotating the textual references that justify the code assignment. Arguably, a higher IAA would have resulted in more top metrics in the CodiEsp-Explainability subtask.

These two limitations, gold standard size and consistency in the text evidence annotation, should act to spur future research groups to extend the CodiEsp corpus. Most participants(78.9%) reported interest in a second CodiEsp edition. Additionally, annotation guidelines are already publicly available, together with the entire CodiEsp corpus, CodiEsp abstracts corpus, and many other resources employed by participants to tune their systems. A second CodiEsp edition could make use of these materials to further promote clinical coding in Spanish.

Closing remarks. The task has been relevant not only in terms of determining the most competitive approaches for this particular data and track, but it also explains how to generate new clinical coding tools for other languages and data collections. The former includes participant systems that are open source and the baseline, all compiled in the CodiEsp webpage [12]. The latter comprises the range of resources described in Section 3: CodiEsp corpus, annotation guidelines, CodiEsp abstracts, terminology mapping chain [DeCS → MeSH → UMLS → ICD-10], machine-translated version of CodiEsp corpus, machine-translated PubMed dump and Codiesp Silver Standard.

Acknowledgements

We acknowledge the Encargo of Plan TL (SEAD) to BSC for funding, and the scientific committee for their valuable comments, their guidance, and their help with the review of the proceedings. Besides, we would also like to thank the organization of CLEFeHealth, Lorraine Goeuriot, Hanna Suominen, and Liadh Kelly. Finally, we would particularly like to thank the team of Bitac, Gloria González, and Toni Mas, who worked with us to create the dataset.

[12] `temu.bsc.es/codiesp/participants-systems/`

A Appendix Title

	all codes					train+dev codes					categories		
	MAP	MAP@30	P	R	F1	MAP	MAP@30	P	R	F1	P	R	F1
TeamX	.284	.277	.123	.448	.192	.334	.326	.123	.522	.199	.162	.538	.249
	.299	.28	.011	.713	.022	.351	.329	.011	.831	.022	.02	.858	.038
	.259	.24	.052	.58	.096	.304	.282	.052	.676	.097	.082	.681	.147
	.265	.24	.004	.858	.009	.311	.282	.004	1	.009	.01	.968	.021
	.065	.053	.011	.713	.022	.08	.066	.011	.831	.022	.02	.858	.038
SWAP	.202	.202	.295	.323	.308	.236	.236	.295	.376	.331	.346	.405	.373
	.013	.013	.026	.041	.032	.015	.014	.026	.048	.034	.04	.064	.049
	.117	.117	.272	.218	.242	.136	.136	.272	.254	.262	.3	.254	.275
	.169	.16	.135	.442	.207	.195	.185	.135	.515	.214	.163	.542	.25
	.013	.013	.04	.03	.034	.022	.022	.099	.033	.049	.096	.082	.089
LIIR	.044	.044	.124	.055	.076	.052	.052	.124	.064	.084	.131	.066	.088
	.011	.011	.066	.029	.041	.013	.013	.066	.034	.045	.085	.043	.057
	.015	.015	.073	.032	.044	.017	.017	.073	.037	.049	.091	.046	.061
	.002	.002	.013	.006	.008	.004	.004	.032	.007	.011	.063	.019	.03
	.006	.006	.04	.018	.024	.006	.006	.04	.021	.027	.051	.025	.033
FLE	.519	.519	.732	.633	.679	.598	.597	.767	.699	.731	.802	.734	.766
	.481	.48	.733	.588	.652	.553	.553	.768	.646	.702	.804	.687	.741
	.501	.501	.74	.604	.665	.576	.576	.775	.665	.716	.807	.714	.758
	.46	.46	.739	.556	.635	.528	.528	.774	.61	.682	.809	.662	.728
IAM	.521	.521	.817	.592	.687	.605	.605	.843	.672	.748	.877	.69	.773
	.511	.511	.789	.591	.676	.605	.605	.789	.689	.736	.837	.682	.752
BCSG	.242	.242	.375	.285	.324	.288	.288	.375	.333	.352	.425	.332	.373
	.259	.259	.407	.287	.337	.306	.306	.407	.335	.367	.461	.333	.387
	.231	.231	.457	.244	.318	.275	.275	.457	.285	.351	.52	.282	.366
	.21	.21	.342	.28	.308	.244	.243	.342	.327	.334	.407	.332	.366
	.128	.128	.235	.215	.225	.149	.148	.235	.25	.243	.284	.268	.276
SINAI	.301	.298	.412	.538	.467	.391	.39	.513	.615	.559	.53	.646	.582
	.314	.311	.443	.544	.488	.414	.413	.551	.621	.584	.567	.642	.602
	.302	.299	.418	.54	.471	.397	.395	.519	.616	.564	.535	.641	.583
	.251	.251	.45	.433	.441	.328	.328	.559	.496	.526	.586	.519	.551
	.291	.288	.402	.528	.456	.377	.376	.51	.604	.553	.513	.624	.564
DCIC-UNS	.097	.097	.385	.09	.146	.159	.159	.772	.099	.175	.768	.123	.212
	.084	.084	.282	.14	.187	.151	.151	.64	.148	.24	.496	.273	.352
	.074	.074	.482	.061	.108	.12	.12	.738	.067	.123	.783	.111	.194
	.078	.074	.128	.261	.172	.184	.184	.436	.271	.334	.294	.355	.322
IMS	.449	.446	.373	.652	.474	.527	.524	.373	.76	.5	.391	.735	.511
	.391	.383	.306	.672	.42	.459	.45	.306	.783	.44	.321	.756	.451
	.389	.378	.299	.682	.416	.461	.452	.306	.785	.441	.316	.767	.448
	.395	.373	.079	.699	.143	.462	.439	.079	.807	.144	.119	.807	.207

	.392	.369	.081	.709	.145	.465	.442	.086	.802	.156	.125	.811	.217
SSN-NLP	.001	.001	.009	.014	.011	.001	.001	.009	.016	.012	.017	.03	.022
	.007	.007	.025	.049	.033	.008	.007	.025	.057	.035	.039	.084	.053
	.004	.004	.014	.019	.016	.005	.005	.014	.022	.017	.023	.036	.028
	0	0	0	.001	.001	0	0	0	.001	.001	.009	.016	.011
MEDIA	.457	.457	.735	.543	.625	.534	.534	.75	.624	.682	.812	.634	.712
	.488	.487	.637	.62	.629	.572	.572	.658	.711	.683	.719	.723	.721
	.462	.461	.526	.63	.574	.545	.545	.549	.721	.623	.597	.748	.664
	.405	.405	.633	.518	.57	.478	.478	.657	.593	.623	.719	.615	.663
Hulat	.115	.115	.866	.066	.123	.138	.138	.935	.071	.132	.889	.074	.137
NLP-UNED	.1	.1	.542	.089	.153	.118	.118	.542	.104	.174	.6	.107	.181
UDC-UA	.368	.367	.587	.511	.546	.435	.434	.587	.595	.591	.609	.581	.595
	.353	.353	.727	.432	.542	.416	.416	.727	.503	.595	.742	.483	.585
	.313	.313	.712	.374	.49	.369	.369	.712	.435	.54	.725	.419	.531
	.359	.358	.399	.582	.473	.417	.416	.399	.678	.502	.406	.651	.5
	.367	.366	.392	.605	.476	.428	.426	.392	.704	.504	.404	.683	.508
CodeICD @IITH	.192	.192	.462	.281	.35	.274	.274	.683	.308	.424	.673	.371	.478
	.18	.186	.106	.281	.154	.266	.266	.584	.308	.403	.478	.373	.419
The Mental Strokers	.517	.517	.551	.638	.591	.604	.603	.551	.743	.633	.624	.736	.676
	.239	.239	.759	.198	.314	.286	.286	.759	.23	.354	.835	.238	.37
ICB-UMA	.482	.46	.004	.858	.009	.567	.542	.004	1	.009	.01	.968	.021
	.471	.449	.004	.858	.009	.554	.529	.004	1	.009	.01	.968	.021
	.455	.43	.002	.897	.005	.536	.509	.004	1	.009	.008	.987	.016
Exeter Chiefs	.076	.075	.117	.188	.144	.088	.088	.117	.219	.152	.132	.223	.166
	.081	.081	.097	.197	.13	.095	.095	.097	.229	.137	.11	.244	.151
	.078	.078	.111	.176	.136	.091	.091	.111	.205	.144	.121	.208	.153
	.082	.082	.111	.201	.143	.097	.096	.111	.234	.151	.126	.239	.165
LSI-UNED	.517	.517	.252	.664	.365	.596	.596	.38	.734	.501	.298	.787	.433
	.493	.493	.252	.666	.366	.571	.571	.366	.737	.489	.299	.792	.434
	.372	.372	.199	.524	.288	.475	.475	.321	.584	.415	.308	.669	.422
	.511	.51	.253	.688	.37	.612	.612	.37	.76	.498	.301	.811	.439
nlp4life	.004	.004	.014	.038	.02	.005	.004	.014	.045	.021	.023	.056	.032
Anuj	.006	.006	.018	.022	.02	.022	.022	.107	.024	.039	.043	.059	.05
	0	0	0	0	0	0	0	0	0	0	0	0	0
	.135	.135	.301	.149	.199	.18	.18	.466	.169	.248	.597	.185	.282
	.505	.505	.741	.621	.676	.596	.596	.741	.724	.732	.78	.696	.735
IXA-AAA	.543	.529	.004	.858	.009	.638	.622	.004	1	.009	.01	.968	.021
	.485	.469	.004	.858	.009	.571	.553	.004	1	.009	.01	.968	.021
	.593	.578	.004	.858	.009	.698	.681	.004	1	.009	.01	.968	.021

Table 9: All metrics of CodiEsp-Diagnostic.

	all codes					train+dev codes					categories		
	MAP	MAP@10	P	R	F1	MAP	MAP@10	P	R	F1	P	R	F1
TeamX	.186	.168	.011	.685	.022	.207	.186	.011	.83	.023	.013	.716	.026
	.182	.168	.042	.415	.077	.202	.187	.042	.503	.078	.048	.454	.087
	.19	.169	.011	.685	.022	.212	.188	.011	.83	.023	.013	.716	.026
	.16	.147	.029	.423	.054	.176	.161	.029	.513	.054	.032	.455	.06
	.166	.147	.004	.825	.008	.183	.161	.004	1	.008	.005	.857	.01
SWAP	.221	.219	.186	.38	.25	.25	.247	.186	.461	.265	.206	.416	.276
	.137	.127	.122	.399	.187	.152	.141	.122	.484	.195	.133	.43	.203
	.141	.14	.155	.323	.209	.154	.153	.155	.392	.222	.16	.339	.218
	.17	.15	.097	.513	.164	.191	.168	.097	.622	.169	.107	.557	.18
	0	0	0	0	0	0	0	0	0	0	0.001	.001	
LIIR	.017	.017	.051	.034	.041	.018	.018	.051	.041	.046	.053	.036	.043
	.007	.007	.015	.01	.012	.008	.008	.015	.012	.014	.017	.012	.014
	.02	.02	.046	.02	.028	.022	.022	.046	.025	.032	.053	.036	.043
	0	0	0	0	0	0	0	0	0	0	.068	.046	.055
	0	0	0	0	0	0	0	0	0	0	.068	.046	.055
FLE	.434	.433	.587	.448	.508	.515	.514	.627	.539	.58	.665	.468	.549
	.433	.432	.587	.446	.507	.513	.512	.626	.537	.578	.665	.465	.548
	.443	.443	.643	.428	.514	.525	.525	.692	.514	.59	.687	.462	.552
	.44	.44	.642	.424	.511	.52	.52	.692	.51	.587	.687	.458	.55
IAM	.426	.426	.659	.373	.476	.496	.496	.659	.452	.536	.761	.431	.55
	.493	.493	.691	.42	.522	.569	.569	.691	.509	.586	.764	.467	.579
SINAI	.28	.275	.367	.452	.405	.33	.326	.393	.548	.457	.441	.492	.465
	.293	.289	.37	.476	.416	.351	.347	.39	.577	.465	.437	.514	.472
	.271	.267	.342	.455	.391	.327	.323	.368	.552	.442	.405	.494	.445
	.25	.245	.343	.422	.378	.298	.293	.373	.512	.432	.402	.456	.427
	.254	.249	.318	.458	.376	.303	.297	.344	.556	.425	.377	.5	.43
MEDIA	.386	.383	.455	.52	.485	.438	.435	.456	.63	.529	.521	.542	.531
	.442	.442	.601	.412	.489	.509	.509	.602	.499	.546	.714	.427	.535
	.404	.402	.501	.503	.502	.457	.454	.502	.608	.55	.586	.525	.554
NLP-UNED	.168	.168	.34	.018	.035	.17	.17	.34	.022	.041	.468	.025	.048
UDC-UA	.22	.22	.389	.268	.318	.251	.251	.389	.326	.354	.428	.293	.348
	.33	.33	.763	.209	.329	.375	.375	.763	.254	.381	.838	.232	.363
	.317	.317	.82	.176	.29	.36	.36	.82	.214	.339	.888	.193	.317
	.269	.269	.379	.322	.348	.307	.307	.379	.39	.385	.423	.359	.388
	.277	.277	.375	.34	.357	.316	.315	.375	.412	.393	.418	.381	.399
The Mental Strokers	.445	.444	.454	.527	.488	.509	.508	.454	.639	.531	.509	.579	.541
	.407	.407	.537	.432	.479	.468	.468	.537	.524	.53	.591	.476	.527
Exeter	.123	.115	.069	.325	.114	.133	.124	.069	.394	.118	.073	.344	.12

Chiefs	.123	.117	.079	.291	.124	.134	.127	.079	.353	.129	.082	.306	.129
	.125	.118	.072	.308	.117	.136	.128	.072	.374	.121	.075	.323	.121
	.121	.119	.106	.228	.145	.132	.129	.106	.276	.154	.1239	.151	
LSI-UNED	.366	.362	.063	.54	.114	.421	.418	.129	.65	.215	.072	.6	.128
	.376	.369	.066	.561	.118	.44	.434	.135	.676	.225	.074	.622	.133
	.351	.346	.056	.48	.101	.403	.399	.073	.582	.129	.06	.5	.106
	.31	.301	.055	.473	.099	.345	.336	.069	.574	.124	.059	.492	.106
	.398	.392	.066	.569	.119	.457	.452	.136	.686	.227	.075	.632	.134
anuj	.069	.069	.216	.025	.045	.124	.124	.415	.03	.057	.235	.028	.05
	.269	.269	.833	.006	.011	.283	.283	.833	.007	.014	.833	.006	.011
	.014	.013	.014	.055	.023	.09	.09	.22	.066	.102	.02	.077	.032
	.413	.413	.602	.396	.478	.474	.474	.602	.48	.534	.665	.439	.529
IXA-AAA	.412	.395	.004	.825	.008	.46	.441	.004	1	.008	.005	.857	.01
	.362	.339	.004	.825	.008	.414	.389	.004	1	.008	.005	.857	.01
	.425	.401	.004	.825	.008	.481	.455	.004	1	.008	.005	.857	.01

Table 10: All metrics of CodiEsp-Procedure.

	all codes			**train+dev codes**		
	P	R	F1	P	R	F1
FLE	.669	.562	.611	.704	.634	.667
	.667	.527	.589	.702	.592	.642
	.687	.537	.603	.725	.604	.659
	.685	.505	.581	.722	.566	.635
IAM	.75	.515	.611	.77	.594	.67
	.732	.524	.611	.732	.616	.669
SINAI	.33	.425	.371	.396	.493	.439
	.36	.447	.399	.428	.517	.469
	.323	.42	.365	.386	.485	.43
	.337	.346	.342	.403	.402	.402
	.313	.421	.359	.382	.49	.429
UCD-UA	.507	.42	.46	.507	.494	.501
	.678	.352	.463	.678	.414	.514
	.671	.303	.418	.671	.357	.466
	.359	.472	.408	.359	.556	.436
	.354	.492	.412	.354	.579	.439
The Mental Strokers	.534	.478	.505	.534	.562	.548
LSI-UNED	.268	.414	.326	.351	.465	.4
	.397	.413	.405	.443	.464	.453
	.508	.406	.451	.537	.457	.494
Anuj	.362	.094	.15	.491	.107	.175
	.572	.456	.507	.572	.536	.553
IXA-AAA	.043	.318	.075	.043	.374	.076
	.144	.301	.195	.144	.354	.205

	.288	.278	.283	.288	.327	.306

Table 11: All metrics of CodiEsp-Explainability.

References

[1] Aitor Gonzalez Agirre et al. "Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track". In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 1–10.

[2] Katherine G Akers. "New journals for publishing medical case reports". In: *Journal of the Medical Library Association: JMLA* 104.2 (2016), p. 146.

[3] Mario Almagro et al. "ICD-10 coding based on semantic distance: LSI UNED at CLEF eHealth 2020 Task 1". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[4] Eiji Aramaki et al. "MedNLPDoc: Japanese Shared Task for Clinical NLP". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. 2016, pp. 13–16.

[5] Alberto Blanco, Alicia Pérez, and Arantza Casillas. "IXA-AAA at CLEF eHealth 2020 CodiEsp Automatic classification of medical records with Multi-label Classifiers and Similarity Match Coders". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[6] José Cañete et al. "Spanish Pre-Trained BERT Model and Evaluation Data". In: *to appear in PML4DC at ICLR 2020*. 2020.

[7] Sébastien Cossin and Vianney Jouhet. "IAM at CLEF eHealth 2020: concept annotation in Spanish electronic health records". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[8] Sébastien Cossin et al. "IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates". In: *arXiv preprint arXiv:1807.03674* (2018).

[9] João Costa et al. "Fraunhofer AICOS at CLEF eHealth 2020 Task 1: Clinical Code Extraction From Textual Data Using Fine-Tuned BERT Models". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[10] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: `1810.04805`. URL: `http://arxiv.org/abs/1810.04805`.

[11] Antje Dörendahl et al. "Overview of the CLEF eHealth 2019 Multilingual Information Extraction". In: (2019).

[12] Sedigheh Eslami, Peter Adorjan, and Christoph Meinel. "SehMIC: Semi-hierarchical Multi-label ICD code Classification". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[13] ExeterChiefs. *eda-classification*. `https://github.com/aollagnier/eda_classification`. 2020.

[14] Nuria García-Santa and Kendrick Cetina. "FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[15] Lorraine Goeuriot et al. "CLEF 2017 eHealth evaluation lab overview". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2017, pp. 291–303.

[16] Hulat. *Codiesp-CLEF-2020-eHealth-Task1*. `https://github.com/pqueipo/Codiesp-CLEF-2020-eHealth-Task1`. 2020.

[17] IAM. *IAMsystem*. `https://github.com/scossin/IAMsystem`. 2020.

[18] *IBECS*. `https://ibecs.isciii.es/`. Accessed: 2020-08-26.

[19] ICB-UMA. *CLEF-2020-CodiEsp*. `https://github.com/guilopgar/CLEF-2020-CodiEsp`. 2020.

[20] Iker de la Iglesia et al. "MEDIA team: CLEF-2020 eHealth Task 1: Multilingual Information Extraction - CodiEsp". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[21] IMS. `https://github.com/gmdn`. 2020.

[22] Ander Intxaurrondo. *AbreMES-DB*. Version 2018-12-01. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL). Zenodo, Nov. 2018. DOI: `10.5281/zenodo.2207130`. URL: `https://doi.org/10.5281/zenodo.2207130`.

[23] Ander Intxaurrondo et al. "The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts". In: (2017).

[24] Ander Intxaurrondo et al. "Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results." In: *IberEval@ SEPLN*. 2018, pp. 280–289.

[25] Rishi Vardhan K et al. "Transformers in Semantic Indexing of Clinical Codes". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[26] *LILACS*. `https://lilacs.bvsalud.org/es/`. Accessed: 2020-08-26.

[27] Carolyn E Lipscomb. "Medical subject headings (MeSH)". In: *Bulletin of the Medical Library Association* 88.3 (2000), p. 265.

[28] Guillermo López-García, José M. Jerez, and Francisco J. Veredas. "CB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[29] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[30] Montserrat Marimon et al. "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results." In: *IberLEF@ SEPLN*. 2019, pp. 618–638.

[31] Toni Mas et al. *CodiEsp guidelines*. Version 1. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL). Mar. 2020. DOI: `10.5281/zenodo.3730567`. URL: `https://doi.org/10.5281/zenodo.3730567`.

[32] Elias Moons and Marie-Francine Moens. "Convolutional Attention Models with Post-Processing Heuristics at CLEF eHealth 2020". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[33] nlp4life. *CLEFeHealth2020-multilabel-bert*. `https://github.com/sarahESL/CLEFeHealth2020-multilabel-bert`. 2020.

[34] Giorgio Maria Di Nunzio. "As Simple as Possible: Using the R Tidyverse for Multilingual Information Extraction. IMS UniPD ad CLEF eHealth 2020 Task 1". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[35] Anaïs Ollagnier and Hywel Williams. "Text Augmentation Techniques for Clinical Case Classification". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[36] José M. Perea-Ortega et al. "SINAI at CLEF eHealth 2020: testing different pre-trained word embeddings for clinical coding in Spanish". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[37] John Pestian et al. "A shared task involving multi-label classification of clinical free text". In: *Biological, translational, and clinical language processing*. 2007, pp. 97–104.

[38] Marco Polignano et al. "A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[39] Paula Queipo-Álvarez, Paloma Martínez-Fernández, and Israel González-Carrasco. "Classifying clinical case studies with ICD-10 at Codiesp CLEF eHealth 2020 Task 1-Diagnostics". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[40] Dietrich Rebholz-Schuhmann et al. "CALBC silver standard corpus". In: *Journal of bioinformatics and computational biology* 8.01 (2010), pp. 163–179.

[41] Henning Schäfer and Christoph M. Friedrich. "Multilingual ICD-10 Code Assignment with Transformer Architectures using MIMIC-III Discharge Summaries FHDO Biomedical Computer Science Group (BCSG)". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[42] Felipe Soares and Martin Krallinger. “BSC Participation in the WMT Translation of Biomedical Abstracts”. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. 2019, pp. 175–178.

[43] Mary H Stanfill et al. “A systematic literature review of automated clinical coding and classification systems”. In: *Journal of the American Medical Informatics Association* 17.6 (2010), pp. 646–651.

[44] Pontus Stenetorp et al. “BRAT: a web-based tool for NLP-assisted text annotation”. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 102–107.

[45] SWAP. *CODIESP-10*. `https://github.com/marcopoli/CODIESP-10`. 2020.

[46] Yuki Tagawa et al. “TeamX at CLEF eHealth 2020: ICD Coding with N-gram Encoder and Code-filtering Strategy”. In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings. 2020.

[47] *UMLS - Metathesaurus*. `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html`. Accessed: 2020-08-26.

[48] Marta Villegas et al. “Esfuerzos para fomentar la minerıa de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologıas del lenguaje”. In: *Procesamiento del Lenguaje Natural* 59 (2017), pp. 141–144.