Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text

Salvador Lima-López[1], Eulàlia Farré-Maduell[1], Luis Gasco-Sánchez[1], Jan Rodríguez-Miret[1] and Martin Krallinger[1,*]

[1] Barcelona Supercomputing Center, Spain

*Corresponding author: Krallinger.Martin@gmail.com

Abstract

Systems able to detect and normalize symptom mentions from clinical texts are crucial for healthcare data mining, AI applied to clinical systems, medical analytics and predictive applications. As opposed to other clinical information types, such as diagnoses/diseases, procedures, lab test results or even medications, clinical symptoms can often only be recovered in detail directly from written clinical narratives. Due to the high complexity, variability and difficulty in generating annotated corpora for clinical symptoms, few manually annotated data collections have been constructed so far. Previous efforts typically showed limitations, such as missing entity normalization to controlled vocabularies, use of dictionaries for pre-annotations, lack of multilingual solutions or underspecified annotation guidelines. To address these issues, we proposed the SympTEMIST track at the BioCreative VIII initiative. The SympTEMIST task is part of the BioCreative VIII evaluation initiative. It is structured into the following three sub-tracks: automatic detection of exact mentions of symptoms, normalizing symptoms to their SNOMED CT concept identifiers and an experimental subtask with the aim of promoting entity linking and concept normalization for several languages, namely English, Portuguese, French, Italian and Dutch. From a total of 25 teams, 11 submitted results for at least one of the three sub-tasks. Top scoring teams obtained an F1-score of 0.7477 for the SymptomNER task (with precision of 0.8039 and recall of 0.6988), while the top-performing team for the SymptomNorm task obtained an accuracy of 0.6070. Taking into account the complexity of symptom mentions, which often include long descriptive or nested entities and abbreviations, the obtained results and used datasets can be considered a relevant contribution for future symptom mining approaches from clinical texts. The SympTEMIST Gold Standard is freely available at: https://zenodo.org/doi/10.5281/zenodo.8223653.

Introduction

Clinical text mining and natural language processing strategies are key to systematically extract clinical variables from medical content, required for the generation of predictive modeling of diseases, healthcare data analytics systems or for generating structured data representations, amongst other possibilities. Most of the existing semantic annotation or clinical concept extraction systems were focused mainly on categories such as medications, diseases, or socio-demographic patient characteristics. As opposed to other clinical information types, such as diagnoses, lab test results or even medications, mentions of symptoms, signs and findings can usually only be recovered directly from written clinical narratives.

During standard medical consultations, symptoms are the patients' complaints that take them to the doctor. They are subjective experiences or sensations reported by patients that most often cannot be directly measured by clinical experts. Symptoms are often described by patients in their own words and include concepts like *pain*, *fatigue* or *discomfort*. In contrast, signs are objective and observable aspects that healthcare professionals can measure and describe through physical examinations and diagnostic or lab procedures, making their evaluation easier. They include concepts such as *abdominal tenderness*, *abnormal heart sounds* or *presence of a mass*.

These two terms are often grouped together under the broader concept of clinical finding. Their descriptive nature makes them essential to learn what is exactly happening to a patient, which will lead to the classification and subsequent targeted treatment of diseases, as well as to detect any adverse effects of treatments and medications.

Textual descriptions of symptoms and signs found within Electronic Health Records (EHR) can be very complex, as they show a high degree of variability that is often heightened depending on clinical report type or clinical speciality. Some of the difficulties include: referring to the same concept using very different terms (such as *fever* or *elevated body temperature*), text spans that include multiple clinical concepts (*hemogram, liver enzymes, CMV and EBV antibodies, renal function and stool culture normal*) and long descriptions of procedures' results (*kidney parenchyma with indeterminate hypoechoic focus measuring 1.5cm noted near the renal pelvis*).

Extracting this information from text requires specialized and comprehensive resources created by experts who have a deep clinical understanding. With this in mind, we propose the SympTEMIST shared task at the BioCreative VIII workshop. SympTEMIST, which stands for SYMPtoms, signs and findings TExt MIning Shared Task, challenges participants to create automatic systems for the detection of mentions of symptoms, signs and findings in clinical texts in Spanish and their subsequent normalization to the clinical terminology SNOMED CT. In order to do this, a Gold Standard corpus of 1,000 clinical case reports annotated with this information was made public. Additionally, to encourage the development of this type of technology for languages other than Spanish, automatically created versions of the corpus translated into eight languages were also released. SympTEMIST builds on previous clinical resources and tasks such as CANTEMIST, DisTEMIST, MedProcNER/ProcTEMIST or MEDDOPLACE. (1-4)

This paper presents an overview of the SympTEMIST task and its resources and results. The sections are distributed as follows: *Task Description* describes the task setting, including the different sub-tasks and their evaluation methods; *Corpus and Resources* introduces the SympTEMIST Gold Standard and other released resources such as the annotation guidelines or the symptoms gazetteer. Next, *Task Results* describes different details about the participation in the task, results obtained by participants and their proposed methodologies. Finally, the *Discussion* section closes off with a short overview of the entire paper and some points for future work.

Task Description

Shared Task Description

The SympTEMIST shared task challenges participants to build automatic systems that can extract different aspects of information about symptoms, signs and findings in texts in Spanish. More specifically, the tasks consist of named entity recognition (NER) and entity linking (EL) of symptom mentions. The entity linking task is explored both from a monolingual and multilingual perspective.

Participants had to use the SympTEMIST corpus, a Gold Standard dataset of 1,000 clinical case reports manually annotated by multiple clinical experts with symptoms, signs and findings and normalized to SNOMED CT codes. SympTEMIST is related to the DisTEMIST (5) and MedProcNER/ProcTEMIST (6) corpora as they were all created using the same clinical case documents, making these annotations highly complementary. The participants' predictions were evaluated against the manual annotations done by clinical experts. Each team was allowed to submit up to 5 runs.

Figure 1 provides a more visual overview of the shared task's setting.

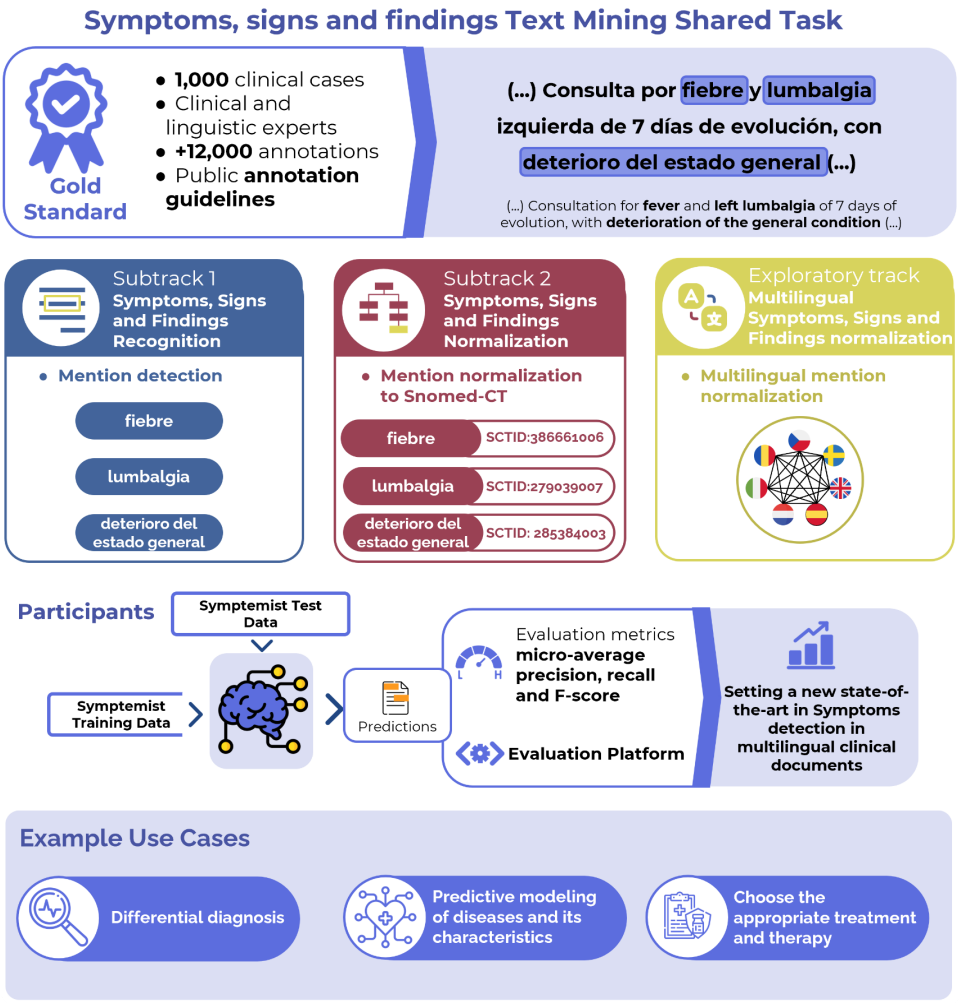

Figure 1. Visual overview of the SympTEMIST shared task setting, originally used for dissemination purposes.

Sub-tasks

SympTEMIST was organized into three different sub-tasks, two monolingual (using Spanish text) and a multilingual one:

- Sub-task 1 (SymptomNER): Symptoms, Signs & Findings Entity Recognition
  This is a named entity recognition (NER) task where participants must automatically detect mentions of symptoms, signs and findings under one single label (SINTOMA, i.e. the Spanish word for symptom) in a test set of 250 documents.

- Sub-task 2 (SymptomNorm): Symptom Normalization & Entity Linking
  This is a normalization or entity linking task where participants are provided a list of entities annotated in the test set and have to return a SNOMED CT code for each of them. Composite mentions were not included in the evaluation. Additionally, only a subset of 450 documents of the training set was released as training data for this sub-task.

- Sub-task 3 (SymptomMultiNorm): Experimental Multilingual Symptom Normalization
  This is an experimental task that uses the automatically annotated Silver Standard multilingual data generated using lexical annotation transfer techniques (7). Participants are provided a list of the mentions transferred from the original Gold Standard in seven languages (Catalan, Dutch, English, French, Italian, Portuguese, Romanian and Swedish) along with each mention's assigned code as training data. For evaluation, participants were given a list of the entities transferred to the test set in Dutch, English, French, Italian and Portuguese and they had to return a single SNOMED CT code for each of them.

Evaluation

The task was evaluated in a two-step sequential process. First, the named entity recognition sub-task was evaluated on its own. Once its evaluation period ended, the complete list of entities annotated in the Gold Standard test set was released (including its automatically generated multilingual transfer) so that participants could generate normalization predictions for all of them. This allowed participants to develop and fine-tune their normalization models on the entire set of recognized entities. This setting is different from the one used for previous tasks like DisTEMIST (2) where the evaluation for both tasks was done in an end-to-end fashion and participants depended on the results of their NER system for the EL sub-task.

NER systems from the first sub-task were evaluated using exact-match micro-averaged precision, recall and F1-score. In addition to this strict evaluation, more relaxed evaluation metrics were also provided to participants in the form of overlapping precision, recall and F1-score. In this variation, predictions that overlap with a Gold Standard mention are considered correct. The two EL sub-tasks were evaluated using accuracy, understood as the percentage of correct mentions out of the total.

Finally, it is worth noting that each of the test sets for every language in Sub-task 3 had a different size. Since the Silver Standards were automatically created, a different number of mentions was correctly transferred in each language. This means that the accuracy results between languages (and between Gold and Silver standards) are not comparable, but only indicative. This was an exploratory novel sub-task to foster the development of multilingual

systems. It was motivated by the pressing need to promote the development of clinical NLP systems, and in particular for clinical concept normalization or entity linking scenarios beyond English and across multiple languages. It was in line with the tasks and objectives for instance of large European projects like DataTools4Heart (https://www.datatools4heart.eu) or AI4HF (https://www.ai4hf.com). The multilingual Silver Standard is described in the *Corpus and Resources* section.

Corpus and Resources

SympTEMIST Gold Standard

The SympTEMIST Gold Standard corpus is a collection of 1,000 clinical case reports in Spanish covering a range of different medical specialties including, amongst others, cardiology, oncology, rheumatology, odontology, urology or mental health. Clinical experts have manually labeled symptoms, signs and findings mentioned in the corpus, using the BRAT annotation tool, and then mapped them to their corresponding SNOMED CT concept identifiers, with an inter-annotator agreement (IAA) of 79.10%. In total, the corpus contains 12,196 annotations (9,092 in the training set and 3,104 in the test set) under the label SINTOMA (the Spanish word for symptoms).

The same clinical case reports were previously also used for other shared tasks, i.e. DisTEMIST for diseases (2) and MedProcNER/ProcTEMIST for clinical procedures (3). Figure 2 shows an example of the annotations provided in the SympTEMIST corpus.

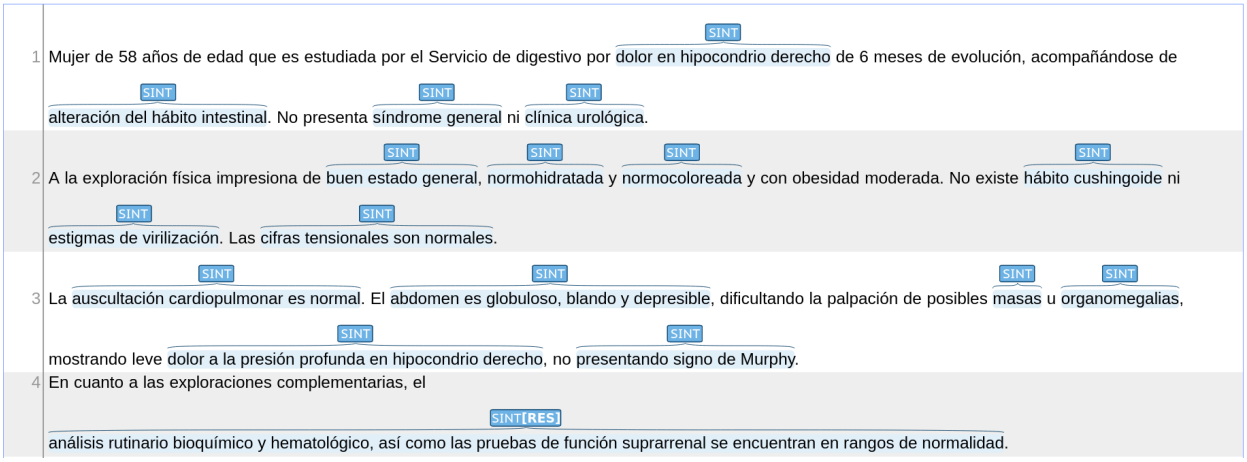

Figure 2. Example of an annotated document from the SympTEMIST corpus.

Translation: "A 58-year-old woman was seen by the gastroenterology department due to *pain in the right hypochondrium* of 6 months of evolution, accompanied by *alteration of the intestinal habitus*. She does not present *general syndrome* or *urological symptoms*. Physical examination showed *good general condition*, *normohydrated*, *normal color* and moderate obesity. There is no *cushingoid habitus* or *virilization stigmata*. The *blood pressure figures are normal*. *Cardiopulmonary auscultation is normal*. The *abdomen is globular, soft and depressible*, making palpation of possible *masses* or *organomegaly* difficult, showing *slight pain on deep pressure in the right hypochondrium*, with no *Murphy's sign*. As for the complementary explorations, the

routine biochemical and hematological analysis, as well as the adrenal function tests were within normal ranges."

The SympTEMIST Gold Standard corpus is publicly available on Zenodo[1].

Additional Resources
As with previous shared tasks, a series of additional resources were released to complement the corpus.

- SympTEMIST Annotation Guidelines:

  The SympTEMIST guidelines describe how to label mentions of symptoms, signs and findings in medical documents in Spanish, as well as how to map or associate them to their corresponding SNOMED CT codes. These guidelines have been meticulously crafted by clinical experts and have undergone iterative refinement through multiple rounds of parallel annotation.

  The first version of the guidelines includes 21 pages and a total of 35 rules divided into different types (general, positive, negative and special). Each rule has an unique identifier, description and definition as well as illustrative examples. Additionally, the guidelines provide an insight into the task's significance and practical applications. They offer fundamental information about the task and the annotation process, delineate different procedure types, draw comparisons with related clinical entity types, and offer guidance and resources for the individuals responsible for annotation. The SympTEMIST Annotation Guidelines are freely available on Zenodo[2].

- Symptoms, signs and findings gazetteer:

  The SympTEMIST gazetteer was created using the Spanish version of SNOMED CT published on October 31, 2022, which includes more than 300,000 concepts. In the construction of the gazetteer, the concepts of the *finding*, *disorder* and *morphologic abnormality* branches were chosen, considering additional sub-hierarchies considering the branches containing the codes assigned by the experts in the corpus. During this process, codes present in the test set were not included, therefore, mentions in the test set that were not in the gazetteer were not taken into account in the evaluation. The gazetteer includes a set of 164,817 lexical entries, comprising 121,713 unique codes distributed in 9 different SNOMED-CT hierarchies. The SympTEMIST gazetteer can be found at Zenodo[3].

- Multilingual Silver Standard:

  The SympTEMIST corpus has been published in a Silver Standard multilingual version to foster the generation and experimental evaluation of multilingual symptom extraction

[1] https://zenodo.org/doi/10.5281/zenodo.8223653
[2] https://zenodo.org/doi/10.5281/zenodo.8246439
[3] https://doi.org/10.5281/zenodo.8413866

and normalization systems. The corpus has been provided in 5 languages (English, French, Italian, Dutch and Portuguese) using a subset of 350 documents from the normalized training set and the entire test set. The process was carried out following a methodology similar to previous work (2-3, 7).

First, an automatic translation of the original documents into the languages of interest was carried out using high-quality commercial machine translation systems. Then, symptom mentions were translated, without incorporating their context, into each of the languages considered. Finally, a transfer of the translated mentions into the documents was made using lexical annotation transfer technologies. In the transfer process, the Snomed-CT code associated with each mention was also included.

The quantity of mentions transferred from the original corpus fluctuated depending on the quality of the machine translation system used and the linguistic similarity between the languages. Table 1 shows the number of mentions in the 350 documents considered in Spanish, and the mentions transferred to each of the languages of the multilingual corpus.

Language	Train set size	Test set size
Spanish (ES)	3484	3104
English (EN)	2003	1600
French (FR)	1802	1425
Italian (IT)	1924	1544
Dutch (NL)	1570	1249
Portuguese (PT)	1686	1521

Table 1. Number of mentions transferred to each of the multilingual Silver Standards' test sets. The Spanish test set corresponds to the Gold Standard, while the others are the languages corresponding to the generated Silver Standards used to evaluate the third sub-task.

- Spanish Silver Standard: This consists of a large collection of over 15 thousand clinical case reports which were publicly released as a background set. Participants then created automatic symptom mention predictions over this background set. Their predictions are released separately as well as in a harmonized version where only matching entities between different sets of predictions are kept using a majority voting mechanism.

Task Results

Participation Overview

A total of 25 teams from 19 different countries registered for the task. Out of them, 11 teams submitted their system predictions for at least one sub-task. In terms of sub-task participation, 10 teams participated in the first sub-task (NER), 7 participated in the second one (EL) and only 2

participated in the third one (multilingual entity linking). All in all, 88 runs were submitted for the entire shared task (42 for NER, 23 for EL and 23 for the multilingual sub-task).

Table 2 shows an overview of the participant teams and their background.

Team's Name	Affiliation	A/I	Country	Sub-tasks	Reference
BIT.UA	IEETA, University of Aveiro	A	Portugal	1, 2, 3	8
BounNLP	Bogazici University	A	Turkey	2	9
FRE	Fujitsu Research of Europe	I	Spain	1	10
Fusion@SU	Sofia University St. Kliment Ohridski	A	Bulgaria	1, 2	11
HIBA	Hospital Italiano de Buenos Aires	A	Argentina	1, 2	12
HPI-DHC	Hasso Plattner Institute, University of Potsdam	A	Germany	1, 2	13
ICB	Universidad de Málaga	A	Spain	1, 2	14
IIC/UC3M	IIC/UC3M	A	Spain	1	15
iML	Carleton University	A	Canada	1	16
PICUS Lab	University of Naples Federico II	A	Italy	1, 2, 3	17
Team-Dm	Freelance	A	Greece	1	-

Table 2. Participant teams in the SympTEMIST task. In the "A/I" column, "A" stands for "academy" and "I" for "industry".

Submission Results

The complete results for the entity recognition, linking and multilingual normalization are shown in Tables 3, 4 and 5, respectively. The top-scoring results for each sub-task were:

Symptoms, Signs & Findings Entity Recognition sub-task. The ICB team obtained the highest strict F1-score (0.7477), as well as the second best F1-score (0.7459), using an ensemble of multiple transformer-based systems. Notably, teams BIT.UA and Fusion@SU also obtained very good results (0.7369 and 0.7324 strict F-1, respectively) by making use of different types of data augmentation. These systems worked especially well in the relaxed (overlapping) evaluation setting. Another noteworthy contribution is that of HPI-DHC team, who obtained a strict F1-score of 0.7363 using the SpanMarker framework[4] (18) and incorporating document-level context to the test set before inference.

Team's Name	Run name	P	R	F1	o_P	o_R	o_F1
ICB	icb-uma-ensemble	**0.8039**	0.6988	**0.7477**	**0.9155**	0.7957	0.8514
ICB	icb-uma-ensemble2	0.7895	0.7068	0.7459	0.9072	0.8122	0.8570
BIT.UA	4-system-all-random	0.7469	0.7271	0.7369	0.8786	0.8553	0.8668
BIT.UA	1-system-all	0.7473	0.7258	0.7364	0.8816	0.8563	**0.8688**
HPI-DHC	2-1e5_last_withcontext	0.7700	0.7049	0.7363	0.8757	0.8009	0.8366
BIT.UA	5-system-all-unk	0.7426	0.7287	0.7356	0.8759	0.8595	0.8676

[4] https://github.com/tomaarsen/SpanMarkerNER

BIT.UA	3-system-top5	0.7411	0.7258	0.7334	0.8757	0.8576	0.8665
Fusion@SU	2-augmented-plantl-roberta-bilstm-crf	0.7393	0.7255	0.7324	0.8766	0.8602	0.8683
HPI-DHC	1-1e5_best_withcontext	0.7667	0.6995	0.7299	0.8762	0.7961	0.8342
BIT.UA	2-system-best	0.7315	0.7274	0.7294	0.8675	0.8628	0.8651
ICB	icb-uma-BioBSC	0.7435	0.7133	0.7208	0.8838	0.8479	0.8655
FRE	2-roberta_ssw	0.7154	**0.7403**	0.7277	0.8487	**0.8782**	0.8632
FRE	1-roberta	0.7231	0.7303	0.7267	0.8616	0.8702	0.8658
HPI-DHC	3-5e5_best_wd_withcontext	0.7586	0.6956	0.7257	0.8781	0.8051	0.8400
Fusion@SU	1-augmented-plantl-roberta-crf	0.7324	0.7178	0.7250	0.8702	0.8528	0.8614
Fusion@SU	3-augmented-symptemist-roberta-crf	0.7149	0.7207	0.7178	0.8603	0.8673	0.8638
ICB	icb-uma-RobertaBioMedical	0.7287	0.7036	0.7159	0.8732	0.8431	0.8579
HIBA	1-model-gz4q4j5o-roberta-clinical-es	0.7276	0.6988	0.7129	0.8742	0.8396	0.8565
Fusion@SU	4-augmented-clin-x-es-roberta-crf	0.7245	0.6991	0.7116	0.8748	0.8441	0.8592
Fusion@SU	5-clin-x-es-roberta-crf	0.7177	0.7026	0.7101	0.8651	0.8470	0.8559
HIBA	4-model-d8km98q7-roberta-clinical-es	0.7291	0.6833	0.7055	0.8910	0.8351	0.8621
HIBA	2-model-ee52hfpj-roberta-clinical-es-retrained	0.7172	0.6936	0.7052	0.8728	0.8441	0.8582
HIBA	3-model-1dcbbqyp-roberta-clinical-es	0.7185	0.6827	0.7001	0.8705	0.827	0.8482
ICB	icb-uma-XLMR-Galen	0.7137	0.6817	0.6937	0.8637	0.8251	0.8440
HPI-DHC	5-1e5_last_nocontext	0.7675	0.6189	0.6852	0.8909	0.7184	0.7954
HPI-DHC	4-1e5_best_nocontext	0.7673	0.6108	0.6802	0.8932	0.7110	0.7917
iML	4-systemBIO	0.6864	0.6205	0.6518	0.8578	0.7755	0.8146
iML	2-systemAll	0.6839	0.6015	0.6400	0.8403	0.7390	0.7864
iML	1-systemMix	0.6821	0.5999	0.6383	0.8429	0.7413	0.7888
iML	3-systemEHR	0.6951	0.5715	0.6273	0.8985	0.7387	0.8108
PICUS Lab	1-NERresult	0.6096	0.5886	0.5989	0.7951	0.7677	0.7812
PICUS Lab	2-NERresult_postprocessed	0.6083	0.5854	0.5966	0.7944	0.7645	0.7792
PICUS Lab	3-NERresult_postprocessed	0.6083	0.5783	0.5929	0.7950	0.7558	0.7749
PICUS Lab	4-NERresult_postprocessed	0.6106	0.5593	0.5838	0.7977	0.7307	0.7627
PICUS Lab	5-NERresult_postprocessed	0.6119	0.5348	0.5707	0.7987	0.6981	0.7451
iML	5-systemXLM	0.6225	0.5026	0.5561	0.8631	0.6968	0.7711
IIC/UC3M	1_rigodr0	0.5076	0.5048	0.5062	0.8082	0.8038	0.8060
IIC/UC3M	3_rigobs16	0.4615	0.4675	0.4645	0.7847	0.7948	0.7897
IIC/UC3M	0_rigoadapt	0.4578	0.4265	0.4416	0.8167	0.7610	0.7879
IIC/UC3M	2_rigodr005	0.4271	0.4343	0.4307	0.7823	0.7954	0.7888
IIC/UC3M	4_rigodr005notest	0.3342	0.3389	0.3365	0.7999	0.8112	0.8055
Team-Dm	symptemistNER*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3. Complete results for all runs submitted to Sub-task 1 (Symptoms, Signs & Findings Entity Recognition), ordered by F1-score. “P”, “R” and “F1” stand for strict precision, recall and

F1-score; “o_P”, “o_R” and “o_F1” stand for overlapping precision, recall and F1-score. Submissions marked with an asterisk (*) had formatting and submission issues.

Symptom Normalization & Entity Linking sub-task. The highest accuracy score was obtained by the HPI-DHC team, who used the xMEN framework[5] (19) to generate candidates with an ensemble of a TF-IDF vectorizer and cross-lingual SapBERT over Spanish and English aliases for all concepts in the SympTEMIST gazetteer. On top of that, they train a BERT-based cross-encoder for re-ranking candidates. Their best run uses the framework’s default configuration with the specification that predictions have to be made for each entity. There is a tie for the second best accuracy (0.589) between Fusion@SU and BIT.UA teams. Both of them use a multilingual SapBERT model with some differences in their approaches: Fusion use a big knowledge base created by augmenting the SympTEMIST gazetteer with synonyms from UMLS and character shifting; BIT.UA, in turn, calculate SNOMED embeddings using the provided resources and use cosine similarity to identify the code that best matches each term. They also experiment with the code acceptance threshold, amongst other variables.

Team's Name	Run name	Accuracy
HPI-DHC	2-xmen_no_nil	**0.6070**
BIT.UA	1-text_snomed_065_t	0.5890
Fusion@SU	3-xlmr-sap-bert-large-train-gazetteer-umls	0.5890
Fusion@SU	1-xlmr-sap-bert-large-train+gazetteer	0.5876
Fusion@SU	4-xlmr-sap-bert-large-train-gazetteer-sliding-window	0.5869
BIT.UA	3-text_snomed_0_t	0.5859
HPI-DHC	1-xmen	0.5848
BIT.UA	5-text_snomed_065_t_base	0.5778
ICB	symptemist_icb-uma_subtask2_20231017	0.5766
BIT.UA	2-snomed-text_065_t	0.5659
Fusion@SU	2-xlmr-sap-bert-large-train-augmented-gazetteer	0.5652
HPI-DHC	4-xmen_no_nil_02	0.5321
HPI-DHC	3-xmen_02	0.5265
PICUS Lab	1-Subtask2Sapbert	0.4816
BounNLP	1-threshold_075_075	0.4721
BounNLP	3-threshold_08_075	0.4721
BounNLP	4-threshold_08_075	0.4640
BounNLP	2-threshold_075_075	0.4636
PICUS Lab	2-Subtask2Roberta	0.4060
BIT.UA	4-text_snomed_1_t	0.4032
PICUS Lab	3-7R1S	0.4032
HIBA	1-HIBA-symptemist-subtask2	0.2785
Fusion@SU	5-xlmr-sap-bert-large-train-gazetteer-snomed	0.0169

[5] https://github.com/hpi-dhc/xmen

Table 4. Results for all runs submitted to Sub-task 2 (Symptom Normalization & Entity Linking).

Experimental Multilingual Symptom Normalization sub-task. Only one team (BIT.UA) participated in all five proposed languages. Their submission builds on their methodology for the monolingual normalization sub-task, incorporating into their pipeline machine translation models from Helsinki-NLP (20) to translate the provided texts in each language back to Spanish. They achieve an accuracy of at least 0.55 for all languages. The other participant in this sub-task, PICUS Lab, participated only in the Italian test set using the same approach proposed for the monolingual normalization sub-task (question-answering model trained from a multilingual SapBERT model plus a lookup mechanism) but with Italian data. Their results range from 0.40 to 0.48 accuracy.

Team's Name	Language	Run	Accuracy
BIT.UA	EN	1-en	**0.725**
BIT.UA	EN	2-en	0.725
BIT.UA	EN	4-en	0.7137
BIT.UA	EN	3-en	0.5265
BIT.UA	FR	4-fr	**0.5733**
BIT.UA	FR	1-fr	0.5726
BIT.UA	FR	2-fr	0.5726
BIT.UA	FR	3-fr	0.3944
BIT.UA	IT	2-it	**0.6703**
BIT.UA	IT	1-it	0.6697
BIT.UA	IT	4-it	0.6626
PICUS Lab	IT	2-ITSapbert	0.5421
PICUS Lab	IT	1-ITRoberta	0.5084
BIT.UA	IT	3-it	0.5065
PICUS Lab	IT	3-7R1SITA	0.2319
BIT.UA	NL	1-nl	**0.6397**
BIT.UA	NL	2-nl	0.6389
BIT.UA	NL	4-nl	0.6317
BIT.UA	NL	3-nl	0.4764
BIT.UA	PT	1-pt	**0.5575**
BIT.UA	PT	2-pt	0.5569
BIT.UA	PT	4-pt	0.5542
BIT.UA	PT	3-pt	0.3254

Table 5. Results of Sub-task 3 (Experimental Multilingual Symptom Normalization).

Methodologies

In general, most teams used some sort of transformer-based approach with some exceptions.

These are the methodologies used by each team:

- **Team BIT.UA**

  For Sub-task 1 (NER), this team proposes a transformer-based solution with masked CRF and data augmentation (specifically, random and unknown token augmentation). Additionally, before creating the final set of predictions, they use an ensemble of different systems. Their approach achieves the third-best strict F1-score (0.7364) and the best overlapping F1-score (0.8688).

  For Sub-task 2 (Entity Linking), they explore a multilingual SapBERT model (21) to calculate embeddings using the SNOMED CT gazetteer. Then, they calculate cosine similarity to identify the code that best matches each term. They also explore using exact matching over the training data followed by the SNOMED CT gazetteer. Their experiments include varying the threshold of acceptance of each code, as well as the order of the dictionaries and the models. This methodology earns them the second-best accuracy value (0.5890).

  For Sub-task 3 (Experimental Multilingual Normalization), they use a pretty similar methodology to the one used in the previous sub-task to participate in all five languages. However, they also integrate into their pipeline machine translation models from Helsinki NLP (20) to translate the texts into Spanish as a first step. Their approach achieves an accuracy of at least 0.55 for all languages.

- **Team BounNLP**

  For Sub-task 2 (Entity Linking), this team attempts to cluster similar symptoms within the same document and then labels the created groups with the same ID if the word embeddings are in the same cluster. They use DBScan as a clustering algorithm (22) and optimize its epsilon value using the silhouette method. Then, a combined dictionary and unsupervised learning approach and Jaro Winkler distance to the BERT Embeddings were used. The pre-trained model they use is “dccuchile/bert-base-spanish-wwm-cased” (23). Their best run achieves an accuracy of 0.4721.

- **Team FRE**

  For Sub-task 1 (NER), this team uses the pre-trained model “PlanTL-GOB-ES/roberta-base-biomedical-clinical-es” (24) fine-tuned using the SympTEMIST training data. Additionally, they incorporate sub-subword information into the model embeddings. Their approach obtains an F1-score of 0.740, as well as the best strict and overlapping recall values (0.7403 and 0.8782, respectively).

- **Team Fusion@SU**

  For Sub-task 1 (NER), this team uses both a RoBERTa model (“PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model” (24)) as well as

CLIN-X-ES model. They fine-tune these pre-trained models using the SympTEMIST training data and apply data augmentation (synonym replacement using terms obtained from the UMLS metathesaurus (25)) on 80% of the data. This methodology earns them an F1-score of 0.732 in the strict setting, as well as the second best overlapping F1-score (0.8683).

For Sub-task 2 (Entity Linking), they make use of a pre-trained SapBERT model ("cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR-large" (21)). As a reference knowledge base, they use the SympTEMIST training data and gazetteer, which they augment for some of the runs using synonyms obtained from UMLS and introducing up to 5 examples for each gazetteer by using character-level perturbations (randomly adding or deleting characters). Their best run achieves the second-best accuracy value (0.5890).

- **Team HIBA**

For Sub-task 1 (NER), this team fine-tuned the pre-trained model "PlanTL-GOB-ES/roberta-base-biomedical-es" (26) on the SympTEMIST training data using different parameters. For one of their runs, they also re-train the model using additional data. Their best run obtains a strict F1-score of 0.7129 and an overlapping F1 of 0.8565.

For Sub-task 2 (Entity Linking), they pre-process the SympTEMIST data using standard NLP tools and try to do a lexical look-up to their own terminological database, which is then mapped to a SNOMED CT code. They achieve an accuracy of 0.2785.

- **Team HPI-DHC**

For Sub-task 1 (NER), this team uses the SpanMarker framework (18) to fine-tune the pre-trained "PlanTL-GOB-ES/roberta-base-biomedical-clinical-es" model using the SympTEMIST training data. Each of their runs use different hyperparameters, with some incorporating document-level context (a feature from SpanMarker) to the test set before running inference with the fine-tuned models. Their approach obtains a strict F1-score of 0.7363 and an overlapping F1 of 0.8366.

For Sub-task 2 (Entity Linking), the team uses a normalization framework called xMEN (19). With it, they generate candidates with an ensemble of a TF-IDF vectorizer and cross-lingual SapBERT over Spanish and English aliases (obtained from the UMLS metathesaurus) for all concepts in the SympTEMIST gazetteer. Then, they train a BERT-based cross-encoder for re-ranking initialized from the pre-trained "PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model".

- **Team ICB**

For Sub-task 1 (NER), this team experimented with different transformers-based solutions. More specifically, they fine-tuned multiple pre-trained models, which they then

ensemble to generate a final set of predictions. Their 5-model-ensemble ("BSC-Bio-Es", "Roberta-Biomedical-Es", "XLMR-Galen", "BETO" and "mBERT-Galen") obtained the absolute best strict F1-score (0.7477) and the best strict and overlapping precision (0.8029 and 0.9150), while their 3-model-ensemble ("BSC-Bio-Es", "Roberta-Biomedical-Es" and "XLMR-Galen") obtains the second best strict F1-score (0.7459) and the second best strict and overlapping precision (0.7895 and 0.9072).

For Sub-task 2 (Entity Linking), they use a pre-trained "SapBERT-XLM-R-large" model and the dense vectors similarity and search library FAISS (27). Their approach earned the team an accuracy of 0.5766.

- **Team IIC/UC3M**

For Sub-task 1 (NER), this team uses a RigoBERTAv2 pre-trained model (28) with different hyperparameter configurations for each run. For one of their runs (run 0), they perform a domain adaptation of the RigoBERTA model using the background data released for the competition. Their approach achieves a strict F1-score of 0.5062.

- **Team iML**

For Sub-task 1 (NER), this team also explores different pre-trained models fine-tuned for the task, both on their own and as ensembles. Their best run is obtained by the "PlanTL-GOB-ES/bsc-bio-es" pre-trained model, which reaches an F1-score of 0.6518.

- **Team PICUS Lab**

For Sub-task 1 (NER), this team uses a similar approach to their participation in the DisTEMIST shared task (29). Additionally, they introduce a similarity-based method in which they calculate the similarity between entities and codes in the gazetteer. Entities with a high similarity to "NO_CODE" are removed. They achieve a strict F1-score of 0.5989.

For Sub-task 2 (Entity Linking), they use a question-answering model fine-tuned from "cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR-large" (as well as "PlanTL-GOB-ES/roberta-base-biomedical-clinical-es") plus a lookup mechanism to detect entities already in the training data. They obtain an accuracy of 0.4816.

For Sub-task 3 (Experimental Multilingual Normalization), this team uses the same approach as for the previous sub-task for the Italian data. Their best run obtains an accuracy of 0.5084.

Discussion

Despite its complexity, SympTEMIST efforts toward the development of NLP systems to automate the capture and normalization of symptoms have shown promising results, with system results close to human annotation quality. However, there are still some complex aspects that

remain as challenges, namely, composite mentions, highly ambiguous mentions and longer text spans, as well as difficulties in differentiating between symptoms and diseases under certain circumstances.

Future directions toward the normalization of symptoms in medical notes should clearly determine the use of the data and the need to further fragment expressions to capture all relevant meaning. Challenges regarding symptom entity linking are related to the underlying annotation workload, as manual mapping is a very time consuming and complex task. The obtained results are competitive but additional training data and resources not only for Spanish but other languages are needed to extract symptoms automatically, including the release and translation of annotation guidelines into multiple languages.

We foresee that for future developments and use of the SympTEMIST datasets some aspects that should also be accounted for are, for instance, gender of the patients associated with the clinical case reports, to assure that systems work equally well for female and male patients. Moreover a more granular analysis of the results of symptom extraction tools depending on the medical specialities covered by each clinical case report would be relevant, also to determine needs with respect to extending the datasets or tasks for specific clinical use case scenarios.

The SympTEMIST dataset and guidelines have already been used for different types of hospital clinical records, as is the case of the CARMEN-I corpus[6] (30) recently released on PhysioNet (31). Thus, it is useful not only to process clinical case report publications but also for different types of hospital medical records, serving as a freely and open access dataset to foster the development of clinical NLP tools overcoming patient data privacy issues.

Together with the Gold Standard SympTEMIST corpus, a silver standard dataset generated by team predictions for a large additional background set of clinical cases has been generated. Additionally, as previously mentioned, the corpus documents used for SympTEMIST have multiple annotation layers for other entities, like diseases, procedures, drugs and medications, chemical compounds and genes and proteins. These resources might be useful to further improve the initial results obtained by participating teams in the current task. How the systems developed using the SympTEMIST dataset would perform when applied to other medical content sources like literature abstracts, clinical trials, patents or healthcare project descriptions (e.g. covered by the MESINESP2 dataset (32)), or even social media (e.g. SocialDisNER effort (33)) remains to be seen.

Beyond the formal quality evaluation of clinical entity recognition and linking systems as done during the SympTEMIST track, public access and release of software tools for clinical NLP remains a pressing need. Promotion of more technical tracks, for instance to implement biomedical NLP component meta-services (34) or interactive tools with experts in the loop (35), could potentially enhance the update and tailor systems to real clinical use cases.

Finally, the results generated by symptom extraction tools from clinical content need to be aligned and integrated into medical application scenarios to prove their practical value. Among the many applications of potential relevance, we see for instance their exploitation for differential diagnosis, analysis and characterization of rare diseases, or serving as features for

[6] https://doi.org/10.13026/bkwd-3j50

predictive modeling systems (e.g. predicting severity of diseases, early diagnosis, adjustment of treatment options).

Funding

This work is supported by the European Union's Horizon Europe Co-ordination & Support Action under Grant Agreement No 101080430 (AI4HF project), Grant Agreement No 101058779 (BIOMATDB project) as well as Grant Agreement No 101057849 (DataTool4Heartproject). We acknowledge the support from the AI4PROFHEALTH Project (PROYECTOS DE I+D+I, PID2020-119266RA-I00) and BARITONE (TED2021-129974B-C22).

References

1. Miranda-Escalada, A., Farré, E., and Krallinger, M. (2020). Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. *IberLEF@ SEPLN*, 303-323.

2. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., ... and Krallinger, M. (2022). Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings.

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

4. Lima-López, S., Farré-Maduell, E., Briva-Iglesias, V., Gasco-Sanchez, L., & Krallinger, M. (2023). MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. *Procesamiento del Lenguaje Natural*, 71, 301-311.

5. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Miranda-Escalada, A., Gasco, L., ... and Paliouras, G. (2022). Overview of BioASQ 2022: the tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. *In International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 337-361). Cham: Springer International Publishing.

6. Nentidis, A., Katsimpras, G., Krithara, A., Lima López, S., Farré-Maduell, E., Gasco, L., ... and Paliouras, G. (2023). Overview of BioASQ 2023: The eleventh BioASQ challenge on large-scale biomedical semantic indexing and question answering. In I*nternational Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 227-250). Cham: Springer Nature Switzerland.

7. Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Estrada, D., Gascó, L., and Krallinger, M. (2022). Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources. *Procesamiento del Lenguaje Natural*, 69, 241-253.

8. Jonker, R. A. A., Almeida, T., Matos, S., and Almeida J. (2023). Team BIT.UA @ BC8 SympTEMIST Track: A Two-Step Pipeline for Discovering and Normalizing Clinical Symptoms in Spanish. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

9. Kavak, B. and Özgür, A. (2023). Symptom normalization using unsupervised learning and text similarity. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

10. Martínez, A. and García-Santa, N. (2023). FRE @ BC8 SympTEMIST track: Named Entity Recognition. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

11. Grazhdanski, G., Vassileva, S., Koychev, I., and Boytcheva, S. (2023). Team Fusion@SU @ BC8 SympTEMIST track: Transformer-based Approach for Symptom Recognition and Linking. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

12. Castaño, J., Franchi, B. C., Benítez, S., Otero, C., and Luna, D. (2023). An exploratory approach to the SympTEMIST challenge. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

13. Borchert, F., Llorca, I., and Schapranow, M. P. (2023). HPI-DHC @ BC8 SympTEMIST Track: Detection and Normalization of Symptom Mentions with SpanMarker and xMEN. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

14. Gallego, F. and Veredas, F. J. (2023). ICB-UMA at BioCreative VIII @ AMIA 2023 Task 2 SYMPTEMIST (Symptom TExt Mining Shared Task). In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

15. Subies, G. G., Jiménez, A. B., and Fernández, P. M. (2023). IIC/UC3M @ BC8 SympTEMIST track: RigoBERTa and Domain Adaptation for SympTEMIST Subtask1. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

16. Shaaban, M. A., Akkasi, A., Khan, A., Komeili, M., and Yaqub, M. (2023). Fine-Tuned Large Language Models for Symptom Recognition from Spanish Clinical Text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

17. Cirillo, M., Moscato, V., and Postiglione, M. (2023). PicusLab @ BC8 SympTEMIST track: Disambiguating Entity Linking Candidates with Question Answering. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

18. Aarsen, T., del Prado Martin, F. M., Suero, D. V., and Oosterhuis, H. SpanMarker for Named Entity Recognition.

19. Borchert, F., Llorca, I., Roller, R., Arnrich, B., and Schapranow, M. P. (2023). xMEN: A Modular Toolkit for Cross-Lingual Medical Entity Normalization. *arXiv preprin*t arXiv:2310.11275.

20. Tiedemann, J., and Thottingal, S. (2020, November). OPUS-MT--Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

21. Liu, F., Vulić, I., Korhonen, A., and Collier, N. (2021). Learning domain-specialised representations for cross-lingual biomedical entity linking. *arXiv preprint* arXiv:2105.14398.

22. Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

23. Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., and Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. *arXiv preprint* arXiv:2308.02976.

24. Carrino, C. P., Armengol-Estapé, J., Gutiérrez-Fandiño, A., Llop-Palao, J., Pàmies, M., Gonzalez-Agirre, A., and Villegas, M. (2021). Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint* arXiv:2109.03570.

25. Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1), D267-D270.

26. Carrino, C. P., Armengol-Estapé, J., Bonet, O. D. G., Gutiérrez-Fandiño, A., Gonzalez-Agirre, A., Krallinger, M., and Villegas, M. (2021). Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *arXiv preprint* arXiv:2109.07765.

27. Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547.

28. Serrano, A. V., Subies, G. G., Zamorano, H. M., Garcia, N. A., Samy, D., Sanchez, D. B., ... and Jimenez, A. B. (2022). RigoBERTa: A State-of-the-Art Language Model For Spanish. *arXiv preprint* arXiv:2205.10233.

29. Moscato, V., Postiglione, M., and Sperlí, G. (2022). Biomedical Spanish Language Models for entity recognition and linking at BioASQ DisTEMIST. In *CEUR Workshop Proceedings* (Vol. 3180, pp. 315-324). CEUR-WS.

30. Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., Murasaki, R. T., Primo-Peña, E., Bojo Canales, C., Paliouras, G., and Krallinger, M. (2021) Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. In *CEUR Workshop Proceedings*.

31. Farre Maduell, E., Lima-Lopez, S., Frid, S. A., Conesa, A., Asensio, E., Lopez-Rueda, A., Arino, H., Calvo, E., Bertran, M. J., Marcos, M. A., Nofre Maiz, M., Tañá Velasco, L., Marti, A., Farreres, R., Pastor, X., Borrat Frigola, X., & Krallinger, M. (2023). CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools (version 1.0). *PhysioNet*. https://doi.org/10.13026/bxrx-y344.

32. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

33. Gasco, L., Estrada-Zavala, D., Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., and Krallinger, M. (2022). The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task* (pp. 182-189).

34. Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C. J., Hsu, C. N., Tsai, R. T., Hung, H. C., Lau, W. W., and Johnson, C. A. (2008). Introducing meta-services for biomedical information extraction. *Genome biology*.

35. Arighi, C. N., Roberts, P. M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-Aryamontri, A., Clematide, S., Gaudet, P., Giglio, M. G., Harrow, I., and Huala, E. (2011) BioCreative III interactive task: an overview. *BMC bioinformatics*.