Computer Methods and Programs in Biomedicine 260 (2025) 108567

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Leveraging Transformers-based models and linked data for deep phenotyping in radiology

Lluís-F. Hurtado [a,c], Luis Marco-Ruiz [b,*], Encarna Segarra [a,c], Maria Jose Castro-Bleda [a,c], Aurelia Bustos-Moreno [d], Maria de la Iglesia-Vayá [e], Juan Francisco Vallalta-Rueda [f]

[a] VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain
[b] Norwegian Centre for E-health Research, University Hospital of North Norway, P.O. Box 35, Tromsø, N-9038, Norway
[c] ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera s/n, València, 46020, Spain
[d] AI Research Unit, MedBravo, Alicante, 03560, Spain
[e] Foundation for the Promotion of the Research in Healthcare and Biomedicine (FISABIO), Avda. de Catalunya, 21, València, 46020, Spain
[f] Laberit, Avda. de Catalunya, 9, València, 46020, Spain

ARTICLE INFO

Keywords:
Natural language processing
Deep learning
Transformers
Linked data
Deep phenotyping
Radiology
Electronic Health records

ABSTRACT

Background and Objective: Despite significant investments in the normalization and the standardization of Electronic Health Records (EHRs), free text is still the rule rather than the exception in clinical notes. The use of free text has implications in data reuse methods used for supporting clinical research since the query mechanisms used in cohort definition and patient matching are mainly based on structured data and clinical terminologies. This study aims to develop a method for the secondary use of clinical text by: (a) using Natural Language Processing (NLP) for tagging clinical notes with biomedical terminology; and (b) designing an ontology that maps and classifies all the identified tags to various terminologies and allows for running phenotyping queries.

Methods and Results: Transformers-based NLP Models, concretely pre-trained RoBERTa language models, were used to process radiology reports and annotate them identifying elements matching UMLS Concept Unique Identifiers (CUIs) definitions. CUIs were mapped into several biomedical ontologies useful for phenotyping (e.g., SNOMED-CT, HPO, ICD-10, FMA, LOINC, and ICPC2, among others) and represented as a lightweight ontology using OWL (Web Ontology Language) constructs. This process resulted in a Linked Knowledge Base (LKB), which allows running expressive queries to retrieve reports that comply with specific criteria using automatic reasoning.

Conclusion: Although phenotyping tools mostly rely on relational databases, the combination of NLP and Linked Data technologies allows us to build scalable knowledge bases using standard ontologies from the Web of data. Our approach enables us to execute a pipeline which input is free text and automatically maps identified entities to a LKB that allows answering phenotyping queries. In this work, we have only used Spanish radiology reports, although it is extensible to other languages for which suitable corpora are available. This is particularly valuable in regional and national systems dealing with large research databases from different registries and cohorts and plays an essential role in the scalability of large data reuse infrastructures that require indexing and governing distributed data sources.

1. Introduction

1.1. Problem

Secondary use of clinical data has seen an important development in the last decade [1]. Both national and international initiatives to enable data reuse across institutions have received significant funding. Examples are: the National Patient-Centered Clinical Research Network, which has accumulated data from 80 million patients and several hundred hospitals since 2013 [2]; the Observational Health Data Sciences and Informatics, which has performed clinical studies at an international level encompassing more than 600 million patient records [3,4]; and the European Health Data and Evidence Network, which aims to build a federated data warehouse of clinical data [5]. These initiatives have played a significant role in designing clinical studies by

* Corresponding author.
E-mail address: luis.marco.ruiz@ehealthresearch.no (L. Marco-Ruiz).

https://doi.org/10.1016/j.cmpb.2024.108567
Received 26 April 2023; Received in revised form 12 November 2024; Accepted 16 December 2024
Available online 3 January 2025

allowing multi-center evaluations and setting the foundations for effective electronic phenotyping. However, Real World Data (RWD) poses significant challenges due to its variability, context-dependent jargon, missing data, and ambiguity [6]. These challenges become exacerbated in clinical notes expressed as free text [7], limiting the possibility of performing effective electronic phenotyping over Electronic Health Records (EHRs) [6,8]. For this reason, the majority of data reuse infrastructures focus on parts of the EHR that are structured (e.g., patient summaries) or normalized cohorts databases, but they do not offer the amount of detail that personalized medicine requires for interpreting highly granular information items (i.e., performing Deep Phenotyping). Deep Phenotyping, understood as "precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described" [9], requires the identification and detailed interpretation of clinical records in the EHR. Some records, such as medications, -Omics data, or laboratory tests, are structured and coded with international terminologies. However, other parts of the EHR, such as radiology reports, are often expressed as free text or semi-structured text and require Natural Language Processing (NLP) to extract the wealth of information in the EHR. Previous studies have used NLP to tackle this task [10–12], but they did not close the data reuse loop by allowing running phenotyping queries over the notes processed with NLP.

In this study, we propose a methodology to advance the reuse of free text radiology reports in Spanish by using NLP transformer technology combined with biomedical ontologies and Linked Data. To that end, our methodology expresses clinical terms identified with NLP as an ontology that allows for running phenotyping queries based on the concept model of SNOMED-CT. This allows identifying granular entities needed for Deep Phenotyping and facilitates the Extraction Transformation and Load (ETL) for secondary use of clinical notes.

The paper is organized as follows. The rest of this section presents the state of the art in NLP, focusing on the biomedical domain, and highlights the novelty of the presented methodology. Section 2 describes the whole system, from the clinical report to the built ontology. Section 3 describes the dataset, use cases, transformer models, and the results of the NLP models. Section 4 presents the ontology-learning pipeline and the ontology built from free text. Finally, Section 5 contains the discussion, the comparison with similar works, and the limitations of this study.

1.2. What is already known

The representations of words or sequences of words (parts of a sentence, whole sentences, paragraphs, or documents) are a fundamental part for the syntactic and semantic text processing. In 2013, Mikolov et al. [13] introduced the concept of word embedding, which represented a major change in how most language processing tasks were approached, and brought significant improvements. Yet, in recent years, the development of the concept of contextual vector models of words (contextual embeddings) meant another critical leap in the results of the area.

In 2018, a new language representation model called BERT (Bidirectional Encoder Representations from Transformers) [14] was introduced. BERT's key technical innovation was applying the bidirectional training of Transformers (an attention mechanism that learns contextual relations between words or sub-words) to language modeling. The Transformer encoder processes the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on its surroundings (both left and right) words. BERT is designed to pre-train deep bidirectional representations from unlabeled text by joint conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of classification tasks. The use of pre-trained language models has been shown to be very effective in improving many NLP tasks [15,16].

Since the BERT proposal, numerous advancements have been made in pre-trained language models, especially in developing models tailored to specific languages and domains [17]. One such model is RoBERTa [18], which builds on BERT's language masking strategy, wherein the system learns to predict intentionally hidden text sections within otherwise unannotated language examples. RoBERTa modifies key hyperparameters in BERT, including removing BERT's next-sentence pre-training objective, and training with much larger mini-batches and learning rates. These adjustments allow RoBERTa to improve on the masked language modeling objective compared with BERT and leads to better downstream task performance.

However, challenges persist when applying models like RoBERTa to non-English languages, particularly Spanish, due to a lack of high-quality, domain-specific Spanish corpora and the linguistic complexity of Spanish itself. Spanish has rich morphological features, such as gender and number inflections, which complicate model training and adaptation. In recent years, collaborative initiatives such as the IberLEF shared tasks [19,20] have promoted the development of resources for Spanish biomedical NLP, but much work remains to achieve the performance levels seen in English-language models. Recent studies have adapted RoBERTa for Spanish biomedical applications. For instance, models used in [21,22] have shown promising results by fine-tuning pre-trained RoBERTa-based models on Spanish clinical data. Notably, these models outperform earlier approaches in some biomedical tasks, demonstrating the effectiveness of RoBERTa's architecture when adapted to Spanish clinical text. Similarly, Carrino et al. [23,24] utilized Spanish RoBERTa-based models for biomedical tasks, achieving better results than other pre-trained models.

The General Language Understanding Evaluation (GLUE) [25] benchmark was proposed in the NLP area. It consisted of a benchmark of nine diverse Natural Language Understanding tasks, an auxiliary dataset for probing models for understanding of specific linguistic phenomena, and an online platform for evaluating and comparing models. GLUE has successfully promoted the development of language representations of general purpose. To facilitate research on language representations in the biomedicine domain, the Biomedical Language Understanding Evaluation (BLUE) [26] benchmark has also been proposed. The benchmark consisted of five tasks with ten datasets covering biomedical and clinical texts with different dataset sizes and difficulties. GLUE and BLUE provide several tasks on which to compare models, and different labeled corpora to do fine-tuning and testing these models. However, it should be noted that all these corpora are in English.

In the biomedical domain, among recent works, PadChest [27] was the first large-scale dataset of chest X-rays and associated reports that implemented deep learning methods for labeling Spanish radiology reports at scale. Reports were labeled with 293 medical entities mapped to the standard Unified Medical Language System (UMLS) terminology and organized as a hierarchical taxonomy. The primary purpose of PadChest was to build a large annotated X-ray dataset to enable the training of supervised deep learning models in medical imaging. Since its publication [28], the dataset has been downloaded 3504 times from BIMCV, from 50 different countries. Its users include private and public health organizations and academic publications making using it [29,30].

Although studies such as the ones aforementioned have used Machine Learning techniques to identify clinical concepts in free text [31], Machine Learning has not been extensively used for mapping other large corpora into multiple biomedical ontologies such as SNOMED-CT, the Human Phenotype Ontology (HPO), or the Foundational Model of Anatomy [10]. In the past, NLP methods outside Machine Learning were used to map clinical free text to terminologies. One of the pioneering works using NLP to map clinical text into a controlled vocabulary was described by Friedman and colleagues [32]. Aronson [33] developed the MetaMap system to map medical text to the UMLS Metathesaurus using a combination of rules and several string-matching

methods. Hammami et al. [34] recently used a rule-based NLP system to tag pathology reports in Italian with ICD-O-M. These methods were successful in mapping free text to terminologies, but the mappings produced were lists of terminology codes not formally expressed as an ontology (i.e., logic model). This translates into the need to directly query for the specific terminology codes (sometimes hundreds or even thousands) rather than using the relationships conveyed in biomedical ontologies to execute expressive queries (subsumption, equivalence, part of, etc.). For example, query for those radiology reports referring to some kind of neuropathy and retrieve all the reports tagged identifying a specific type of neuropathy (Myasthenia gravis, Iatrogenic neuropathy, Lipoma of nerve, etc.).

1.3. What this paper adds

This paper presents a methodology that uses transformer-based methods to process clinical reports in Spanish and express the extracted entities as an ontology compliant with Linked Data Principles. The innovation of this methodology is threefold. First, it leverages transformers technology on radiology reports written in Spanish and analyzes their performance over clinical reports, which include COVID tags. Second, the extracted entities are directly mapped to 12 biomedical terminologies using SNOMED-CT as the reference one. Third, all terminology mappings are expressed as a Linked Knowledge Base (LKB) allowing for using Linked Data technologies to run phenotyping queries.

2. The NLP-LKB system

The Gobierno de Canarias and the Generalitat Valenciana, through the Servicio Canario de la Salud and the Conselleria de Sanitat Universal i Salut Pública, respectively, considered innovation as an essential tool for improving health care. Consequently, the "Big Data Personalized Medicine" (MedP) project was launched [35], processed under the second call of the FID Salud Program, to support clinical decisions aimed at each individual patient with particular attention to chronic pathologies, and also creating a new patient interface supported by Artificial Intelligence. The MedP Project believes that innovation is fundamental in improving medical care. The FID Program is a government tool funded with FEDER pluri-regional funds for the period 2014–2020, and its objective is to promote innovation through the demand for innovative solutions by public administrations.

Within the framework of this project, we have worked on a use case whose main objective was the application of NLP in the domain of clinical reports. In particular, the goal was the knowledge extraction and automatic labeling in various terminologies (including SNOMED-CT coding) from conventional chest radiology reports.

A scheme of the system is shown in Fig. 1. As illustrated, the system processes clinical reports written in Spanish and expresses the extracted entities as an ontology compliant with Linked Data Principles. This figure depicts an example of two reports processed through the NLP stages (see 1), terminology mapping (see 2), and ontology development to create the LKB (see 3). As shown in the figure, the output of the NLP stage (1) is the set of UMLS Concept Unique Identifiers (CUIs), to which the text segments were classified. In stage (2), using the set of CUIs, mappings among several terminologies of interest (e.g., HPO, FMA, ICD10 etc.) are defined. In stage (3), these mappings are processed to define a LKB supporting phenotyping queries over the mapped concepts.

3. The natural language processing module

The problem can be stated as follows: given a clinical report in Spanish, the objective is to automatically obtain the sequence of labels that explains the report in terms of radiographic findings and differential diagnoses. For example, given the report (in Spanish):

Fecha: 22/03/2020

Juicio clínico: valorar evolución radiológica neumonía por covid 19. En el momento actual se visualiza discreto aumento de la condensación pulmonar a nivel del campo medio derecho., Ligero empeoramiento radiológico.

From this free text clinical report, the system automatically extracts the following labels: "COVID-19", "consolidation", and "pneumonia". There is a 1-to-1 correspondence between labels and CUIs, therefore the sequence of CUIs (in the same order that the labels) that explains the clinical report is: C5203670, C0521530, and C0032285.

The task comprises three steps:

1. Preprocessing: Cleaning the report and extraction of the individual sentences.
2. Obtain the sequence of labels for each sentence.
3. The result is the union of the obtained sequence of labels for each sentence.

In order to obtain the sequence of labels for each sentence (step 2), a multi-label classifier is trained. Multi-label classification is a variant of the classification problem where multiple labels may be assigned to each instance. In our case, the instances are the sentences extracted from the clinical reports, and the labels are radiographic findings and differential diagnoses. The description of the set of predefined labels is presented in the next section. The multi-label classifier is a transformer-based model, described in Section 3.2.

3.1. Materials

A new dataset, comprising sentences extracted from biomedical reports in Spanish, was built to train our classification models. Those biomedical reports came from two different corpora, PadChest and BIMCV COVID-19, described below.

PadChest is a publicly labeled large-scale, high-resolution chest X-ray dataset, and associated reports written in Spanish [27]. This dataset includes more than 160,000 images obtained from 67,000 patients. Physicians developed PadChest by manually reviewing and identifying radiological findings from 22,120 unique sentences to the concepts that best identified the finding semantics in UMLS. Expert physicians manually annotated 27% of the reports, and the remaining reports were labeled using a supervised method based on a recurrent neural network with attention mechanisms. This resulted in a corpus made of multi-labeled sentences and reports where each radiological entity is mapped to one UMLS CUI. Radiological entities were: 174 different radiological findings, 19 differential diagnoses, and 104 anatomic locations, organized as a hierarchical taxonomy and mapped onto standard UMLS terminology. A detailed description of the development of PadChest is available at Bustos et al. [27], and it can be downloaded from https://bimcv.cipf.es/bimcv-projects/padchest/.

The BIMCV COVID-19 dataset is a large open multi-institutional dataset that provides the open scientific community with data of clinical-scientific value that will help the early detection and evolution of COVID-19 [28]. It is an annotated dataset that contains chest X-ray images (CR and DX) and computed tomography (CT) imaging of patients with COVID-19 and no COVID-19 patients. Their radiological reports (in Spanish) are also attached, along with their radiological findings (the same 174 labels as PadChest) and locations, pathologies, DICOM metadata, Polymerase chain reaction, Immunoglobulin G, and Immunoglobulin M diagnostic antibody tests. The findings have been mapped onto standard UMLS terminology, covering a broad spectrum of thoracic entities. Two new labels for diagnoses are added, "COVID-19" and "COVID-19 uncertain". In addition, 23 sample images were annotated by expert radiologists to include semantic segmentation of radiological findings. The dataset can be downloaded from http://bimcv.cipf.es/bimcv-projects/bimcv-covid19/, and it is constantly

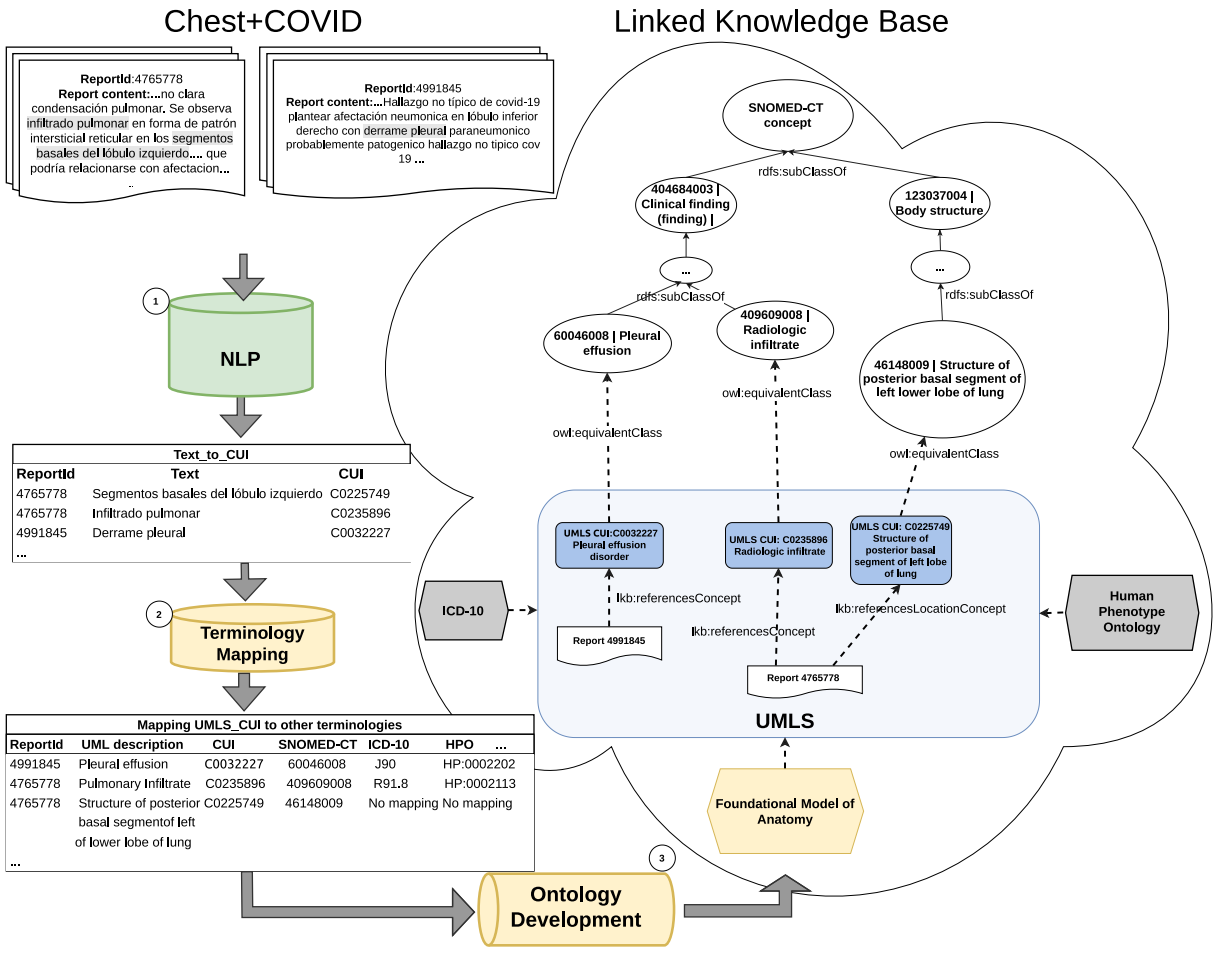

Fig. 1. Stages of the methodology: NLP, terminology mapping, and ontology development.

updated with new samples. At this time (second iteration), BIMCV COVID-19+ dataset comprises 7377 CR, 9463 DX, and 6687 CT studies.

Both publicly available datasets, PadChest and BIMCV COVID-19, were used to build a new dataset, called Chest+COVID, of biomedical reports in Spanish, choosing the manually annotated reports from both datasets for more than 26,000 reports. These reports were preprocessed to extract sentences, resulting in 28,128 sentences after eliminating duplicates. A total number of 299 labels appeared in the dataset (174 radiographic findings, 19 differential diagnoses, and 104 anatomic locations from the PadChest dataset, and the two new labels included in BIMCV COVID-19). The anatomic locations were extracted by using regular expressions, and the remaining 195 labels were extracted with transformers. The Chest+COVID dataset was stratified random sampling and split into training (25,315 sentences), validation (1,406 sentences), and test (1,407 sentences) partitions. Some statistics are shown in Table 1. A histogram of the most frequent labels in each partition is shown in Fig. 2. Specifically, the histogram accounts for the 80 labels (out of 195) with more samples in the training partition.

Table 1
Statistics of the biomedical Chest+COVID dataset: the number of sentences, words, and labels, detailing the total and the average number of words and labels in the sentences. The vocabulary is composed of 5,815 words.

	Sentences	Words		Labels	
		(total)	(average)	(total)	(average)
Training	25,315	263,999	10.43	32,989	1.30
Validation	1,406	14,617	10.40	1,847	1.31
Test	1,407	15,011	10.67	1,849	1.31
Total	28,128	293,627	10.44	36,685	1.30

3.2. Methods

The Chest+COVID dataset was used to fine-tune a transformer based on RoBERTa architecture [18], for the multi-label classification task. In particular, we used the so-called RoBERTa-base-biomedical-clinical-es model [23], trained by the Barcelona Supercomputing Center under the Plan de Impulso de las Tecnologías del Lenguaje of the Spanish Government. The RoBERTa-base-biomedical-clinical-es model is a biomedical pre-trained language model for Spanish. It is ready-to-use only for masked language modeling to perform the Fill Mask task (predicting which words should replace a mask, a gap in a sentence). However, it is intended to be fine-tuned on downstream tasks such as Named Entity Recognition or Text Classification. Our goal is to fine-tune this model for multi-label classification of the sentences extracted from biomedical reports and our set of 195 labels. The selection of the pre-trained RoBERTa-base-biomedical-clinical-es Spanish language model was primarily motivated by its use of the RoBERTa model, which is considered state-of-the-art for many NLP tasks in languages such as English. Furthermore, in the case of Spanish, previous research has compared this model with other language models, such as those discussed in [23,24]. It concluded that using pre-trained models specific to the biomedical domain leads to the best results for various NLP tasks in biomedicine. Several other studies have reached similar conclusions for English [36].

The training dataset for the RoBERTa-base-biomedical-clinical-es model comprises various biomedical corpora in Spanish, compiled from publicly available corpora and trackers, and a clinical dataset from

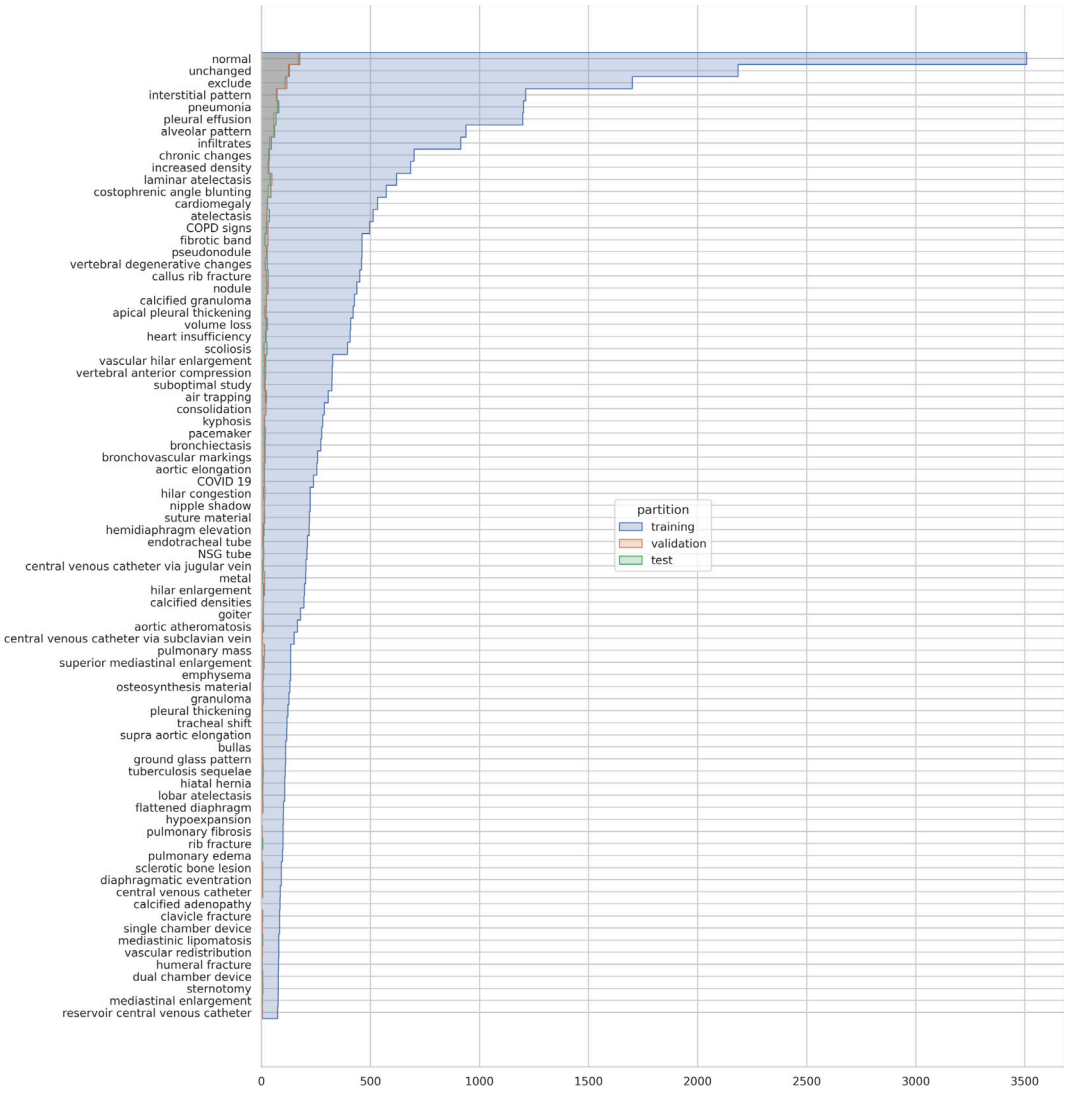

Fig. 2. Histogram of the 80 most frequent labels (out of 195) in training, detailed for each partition of the Chest+COVID dataset.

more than 278,000 documents and clinical notes. In order to obtain a high-quality training dataset while preserving the idiosyncrasies of clinical language, the cleaning pipeline has been applied only to the biomedical dataset, keeping the clinical dataset uncleaned. Essentially, the cleaning operations used are: sentence separation, language detection, filtering of poorly constructed sentences, removal of duplicate content, keep the limits of the original document.

The biomedical corpora were concatenated, and additional global deduplication was applied among the biomedical corpora. Finally, the clinical dataset was concatenated with the clean biomedical dataset, resulting in a medium-sized biomedical-clinical dataset for Spanish composed of more than 1,000 millions tokens. The training partition was tokenized using a byte version of Byte-Pair Encoding used in the original RoBERTa model with a vocabulary size of 52,000 tokens. Then, the masked language model was re-trained at the subword level following the approach used for the base model of RoBERTa with the same hyperparameters proposed in the original work. All the details of the pre-trained RoBERTa model can be found at: https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es.

As stated before, the RoBERTa-base-biomedical-clinical-es model was fine-tuned by using the Chest+COVID dataset to train a multi-label transformer classifier to assign the corresponding set of labels to each sentence extracted from the clinical reports. Specifically, our model consisted of 12 transformer layers, each with 768 hidden units and 12 attention heads. The final layer was a fully connected (dense) layer with a Tanh activation function, allowing the model to perform multi-class, multi-label classification. This setup enables the model to assign multiple labels to each input sentence. In total, the model has $125,978,112$ parameters, allowing for a robust and scalable classification process. The model can be downloaded on our Hugging Face page (at "https://huggingface.co/ELiRF/Chest-COVID/").

3.3. Experimentation and results

The NLP module assigns some labels (out of a set of 195 labels) to each report, but the classification process is done at the sentence level. During inference, the model classifies the report sentences. Sentence-level predictions are then combined to obtain the label sequence of a complete report, keeping in mind that some classes, such as "normal" or "exclude", are exclusive and are eliminated in case the model predicts them together with any other label. For this reason, both the training of the model and the evaluation of its performance were done at the sentence level. In the following subsections, the training strategy of the classification models and the evaluation of their performance on the test partition are discussed.

3.3.1. Experimentation

The classification task addressed in this work is challenging due to the large number of labels and the imbalanced distribution of these

Table 2
Hyperparameters considered in the optimization process.

Hyperparameter	Values considered
hidden_dropout	[0.05, 0.1]
attention_probs_dropout	[0.05, 0.1]
num_train_epochs	[100, 150]
learning_rate	[2e−5, 1e-4]
batch_size	{16, 32}
weight_decay	[0.001, 0.01]
lr_scheduler_type	{constant, linear}

labels. Such imbalance typically causes the model to focus primarily on classes with the highest number of samples, often neglecting those that are rarely encountered during training. To mitigate this issue, we selected the macro F1-score as the objective function for the fine-tuning process. This metric averages the F1 scores of all classes equally, disregarding their sample frequency, thus encouraging balanced performance across classes.

For model fine-tuning, a hyperparameter optimization was performed to identify optimal values that maximize the macro F1-score on the validation set. We conducted a total of ten optimization runs: five with a linear learning rate scheduler and five with a constant learning rate scheduler. The Optuna hyperparameter optimization framework [37] was used for this purpose, as it provides efficient searching and pruning algorithms. Table 2 details the parameters subject to optimization. The other hyperparameters of the base model (such as the number of attention heads per layer, number of layers, feed-forward layer dimensionality, and the dimensionality of Q, K, and V projections) were kept unchanged. Complete details for these parameters can be found in the base model's configuration file at https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es/blob/main/config.json. The loss function used was binary cross-entropy, calculated between the reference labels and the probability predicted by the model for each label.

The results on the validation set during the fine-tuning showed that the runs that used a linear scheduler achieved better results than those that used a constant scheduler. For this reason, we focused the analysis on the five runs with a linear scheduler for the learning rate ("lr_scheduler_type=linear"). The values obtained by these five runs were relatively similar.

The most relevant hyperparameters of each run are shown in Table 3. The selection criteria for the best epoch of each run was the macro F1-score on the validation partition. That is, for each run, the epoch that maximized the result of the macro F1-score on the validation set was chosen. Table 4 shows the macro averaged results of the best epoch of the five runs on the validation set in terms of precision, recall, and F1-score. Run-3 is the model which obtained the maximum value for macro F1-score on the validation set (a value of 0.6996 at epoch 94).

3.3.2. Results

After fine-tuning was completed and the hyperparameters were optimized, the best model was evaluated on the test set. Table 5 presents the performance metrics (precision, recall, and F1-score) of the best run at its optimal epoch. We assess three different averaging methods: micro averaging (assigns equal weight to each sample), macro averaging (assigns equal weight to each label), and weighted averaging (weights each label based on support). The macro averaged measures were obtained by averaging on the 195 labels which appeared in the training partition, while only 151 did appear in the test partition. As stated above, the evaluation is conducted at the sentence level, as both training and ground-truth labeling were performed at this granularity.

The model's macro-averaged results on the test set are consistent with those observed during fine-tuning. As is typical in multi-label classification tasks with substantial class imbalance, the values for macro averaged measures are lower than those for micro or weighted averaged values. It is important to note that macro results are computed by averaging over all classes in the task, regardless of whether they appear in the test partition or are simply hypothesized by the model.

To better understand how class imbalance impacts model performance, we conducted a detailed analysis of the model's performance across different label frequency groups. Fig. 3 illustrates the evolution of micro and macro metrics as a function of label frequency, with labels ordered by their occurrence in the training set. This figure also shows the number of test samples associated with each label.

This analysis reveals a noticeable drop in macro metric performance after the 100 most frequent labels. Consequently, model performance was evaluated using three specific label sets: (1) the 100 most frequent labels in the training set, (2) the remaining, less frequent, labels (95 labels), and (3) labels present in the test set (151 labels). Table 6 presents these results. A breakdown of results by label in terms of true positives, false positives, false negatives, precision, recall, and F1-score is provided in Appendix A.

The model performs best when limited to the 100 most frequent labels in the training set, where these labels are well-represented, leading to stronger predictive outcomes. The results for the 95 less frequent labels, though challenging, are also noteworthy; among these, only 34 labels appear in the test set, yet the model achieves competitive performance, demonstrating its adaptability. Lastly, the most insightful results come from focusing only on labels present in the test set, providing a fairer assessment by excluding labels absent from the test set from the computation formulas. This approach avoids the dilution effect caused by including labels with no test examples, resulting in a clearer view of the model's performance compared to the aggregate results shown in Table 5.

4. Terminology mapping and phenotyping queries

4.1. Methods

4.1.1. Methodology for terminology mapping

The input to the terminology mapping stage is the set of UMLS CUIs assigned to each radiology report by the NLP module. Querying biomedical reports using various coding systems requires mapping the extracted UMLS CUIs to these code systems. In the case study presented, the terminological mapping has been carried out from CUIs to SNOMED-CT, ICD-10, the Human Phenotype Ontology (HPO), ATC, LOINC, ICPC2, MSHSPA (MeSH Spanish version), MDRSPA (MedDRA Spanish version), MedlinePlus, NCI_PI-RADS, NCI_caDSR and Foundational Model of Anatomy (FMA). We reused mappings already available in the UMLS Metathesaurus. Mappings were retrieved by querying the UMLS terminology service to retrieve the mappings available for each CUIs assigned to each report. When mappings were not available for a given UMLS CUI, ad hoc mappings were developed by searching in the online browser of SNOMED International, the definition associated with each CUI, and selecting the best candidate (see Appendix B).

4.1.2. Methodology for linked knowledge base development

Enabling the execution of phenotyping queries at a computational level requires a logic model that allows the analysis of relationships (e.g., equivalence mappings and parent–child relationships) among terminological concepts. Following the methodology described in previous works [38], we developed a SNOMED-CT model expressed in RDF(S). In compliance with ontology design patterns [39], this model provided a logic underpinning (i.e., the main taxonomy) to which concepts from other terminologies were mapped. This implies that the resulting ontology will rely on the organization of SNOMED-CT's polyhierarchy, and concepts from other terminologies will be attached to the correct level of it using equivalence axioms between each concept and its SNOMED-CT equivalent. When a mapping was available, UMLS CUIs were mapped to SNOMED-CT using OWL (Web Ontology Language) axioms. Similarly, all the other concepts from different terminologies

Table 3
Hyperparameters of each run. The rest of the parameters do not vary (num_attention_heads = 12; num_hidden_layers = 12; hidden_act = gelu; hidden_size = 768; intermediate_size = 3072; max_position_embeddings = 514).

Parameter	run-1	run-2	run-3	run-4	run-5
hidden_dropout	0.0943	0.0812	0.0592	0.0843	0.0515
attention_probs_dropout	0.0501	0.0530	0.0936	0.0704	0.0662
num_train_epochs	126	142	139	145	143
learning_rate	9.9078e−5	3.3736e−5	6.1254e−5	6.8333e−5	2.2627e−5
batch_size	32	16	32	32	32
weight_decay	0.0011	0.0057	0.0010	0.0094	0.0034
lr_scheduler_type	linear	linear	linear	linear	linear

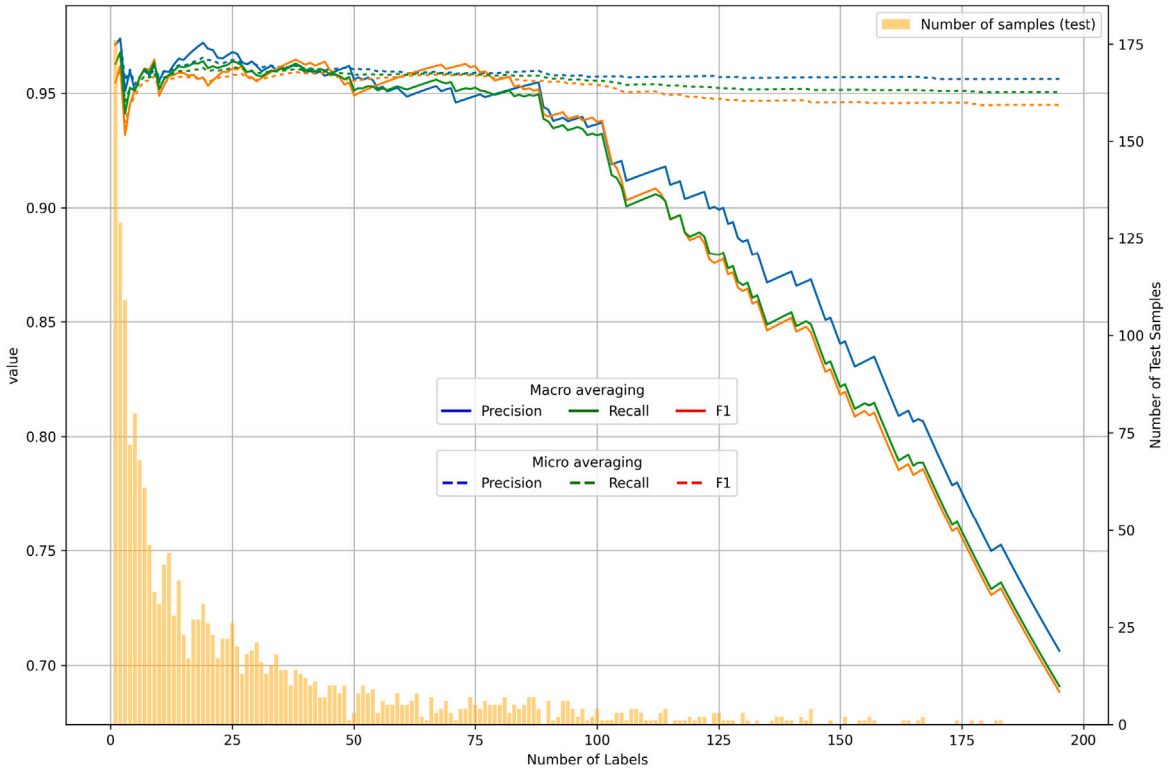

Fig. 3. Evolution of performance metrics (micro and macro averaged) as a function of label frequency. Labels are ordered according to frequency in the training set, showing model performance across high- and low-frequency classes.

Table 4
Macro averaged results of the best epoch of the five runs on the validation set in terms of Precision, Recall, and F1-score. The total number of labels in the validation set is 1,847.

Run	Best Epoch	Precision	Recall	F1-score	Support
run-1	37	0.7006	0.6939	0.6918	1,847
run-2	85	0.7125	0.6993	0.6983	1,847
run-3	94	0.7193	0.6948	**0.6996**	1,847
run-4	45	0.7098	0.6933	0.6949	1,847
run-5	61	0.7065	0.6896	0.6905	1,847

Table 5
Overall results on the test set, showing micro, macro and weighted averages for Precision, Recall, and F1-score. The total number of labels in the test set is 1,849. Macro averaged measures obtained by averaging on the whole set of 195 labels which appeared in the training partition.

	Precision	Recall	F1-score	Support
Micro averaging	0.956	0.945	0.950	1,849
Macro averaging	0.706	0.688	0.691	1,849
Weighted averaging	0.957	0.945	0.949	1,849

were mapped to their corresponding UMLS CUI. This chain of equivalence relationships allowed the triple store DB to reason transitively, determining equivalences among concepts even when they were not directly linked. All the axioms and properties of the model were expressed following Linked Data Principles for scalability and interoperability reasons. Thus, the knowledge graph developed forms a LKB, which allows for reasoning over the relationships and axioms defined [38]. SNOMED-CT was chosen as the central reference ontology (and hierarchical logic model) to structure the LKB because it has extensive coverage of clinical concepts and an underlying machine-understandable logical model, which other terminologies lack. A representation in lightweight semantics (RDF(S)) was chosen over more expressive logics (e.g., OWL-DL) because it is computationally lighter and more compatible with graph databases and triple stores. The LKB was deployed in a triple store DB (Graph DB version 10.1), which supported expressive queries in SPARQL language. Fig. 4 shows the results of the terminology mapping and phenotyping queries executed.

Beyond biomedical ontologies, three other ontologies were used to build the LKB: the Dublin Core (DC), schema.org, and the Minimal Service Model (MSM). The Dublin Core Metadata Initiative (DCMI) (www.dublincore.org/) is an international project currently part of ASIS&T. It focuses on developing best practices and ontologies for the

Table 6
Performance on the test set for different subsets of labels, divided into high-frequency and low-frequency groups. The table presents Precision, Recall, and F1 scores, for micro and macro averaging. The high-frequency label group contains 1,815 labels, while the low-frequency group contains 34 labels, offering a detailed view of model performance across varying label frequencies. The total number of labels in the test set is 1,849. Macro averaged measures for high- and low-frequency labels were obtained by averaging on the whole set of 195 labels which appeared in the training partition. Macro averaged measures for all test labels were obtained by averaging on the reduced set of 151 labels which appeared in the test partition.

Label Set	Averaging Type	Precision	Recall	F1-score	Support
100 Most Frequent in Training	Micro averaging	0.957	0.947	0.952	1,815
	Macro averaging	0.885	0.863	0.866	1,815
Remaining Labels in Training	Micro averaging	0.935	0.853	0.892	34
	Macro averaging	0.349	0.338	0.341	34
Test Labels	Micro averaging	0.958	0.945	0.951	1,849
	Macro averaging	0.912	0.889	0.892	1,849

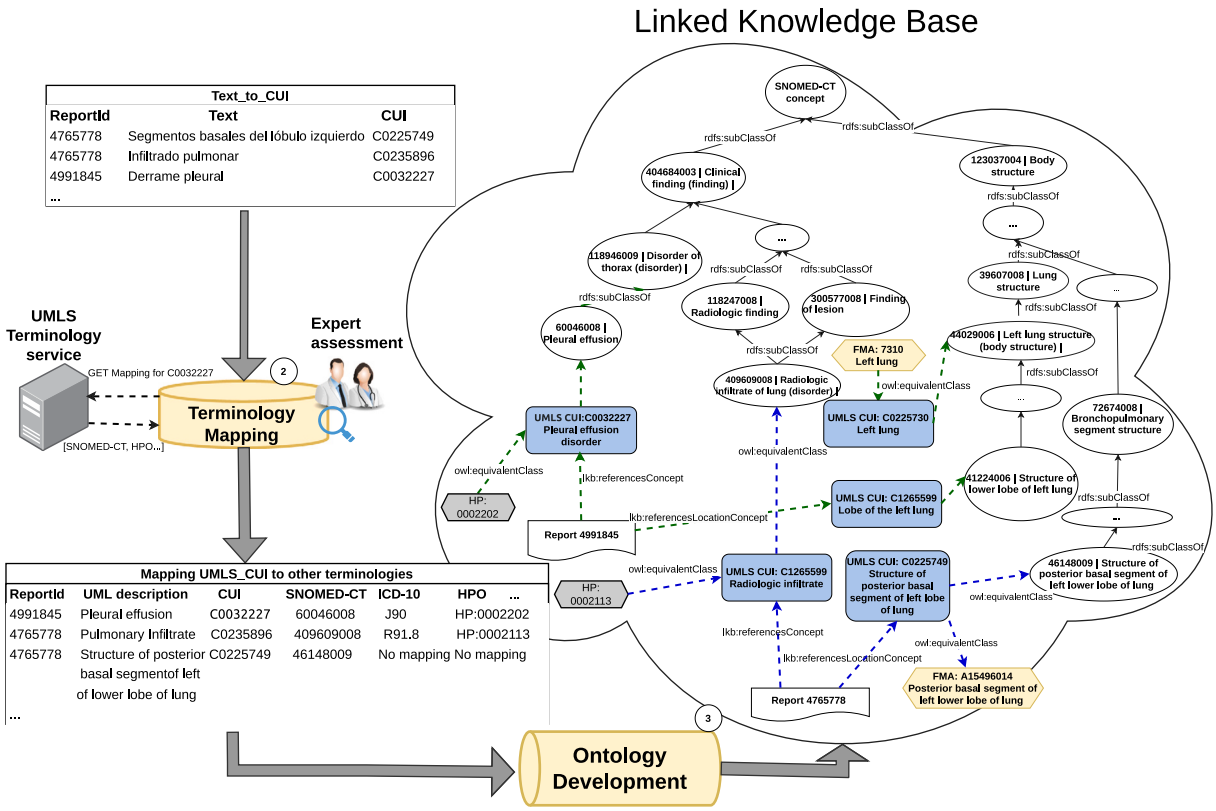

Fig. 4. Terminology mapping and ontology development of the Linked Knowledge Base.

specification of metadata about physical or online resources. Its developments have become the de facto standard for catalogers and open government data publication. Schema.org is a collaborative community promoted by private actors such as Google, Yahoo, Microsoft, and Yandex. Since 2011, schema.org has gained momentum for improving semantic searches on the internet. While the DC provides more abstract ontologies to attach metadata, schema.org has expanded to specific domains such as medicine. In our developments, we prioritized the use of DC because it is more established and aligned with standardization organizations such as W3C. When DC concepts did not suffice to represent the metadata about reports, we used concepts from schema.org. A different ontology used in the design of the LKB is MSM. Its nature differs from the DC and schema.org in that it is not intended to specify resources metadata but to provide a link between Web services and the ontological specification of an LKB. Therefore, its concepts were used to specify what a report is and 'glue' the ontology developed using SNOMED-CT to metadata expressed with DC and schema.org specifying which Web services expose these reports.

4.2. Results

Fig. 4 depicts the stages regarding terminology mapping (see 2) and ontology development to create the LKB (see 3). As explained in previous sections, the output of the NLP stage was the set of UMLS CUIs to which the text segments were classified.

In the second stage (2), CUIs were mapped to SNOMED-CT, ICD-10, Human Phenotype Ontology (HPO), ATC, LOINC, ICPC2, MSHSPA (MeSH Spanish version), MDRSPA (MedDRA Spanish version), MedlinePlus, NCI PI-RADS, NCI caDSR, and Foundational Model of Anatomy (FMA). The release of SNOMED-CT selected was the US version. While the choice of the US version may seem counterintuitive, it was motivated by the fact that the coverage of mappings between UMLS CUIs and concepts from the SNOMED-CT US version is higher than in the Spanish version. For example, at the moment of writing, the CUI C1709576 has a mapping in SNOMED-CT US to "Mass of pleura" (SCTID=540881000124100), but there is no mapping available to the Spanish version (SCTSPA) in UMLS. Note that this choice on the

SNOMED-CT version does not affect NLP performance because NLP, as described in Section 3, is only mapped to CUIs. The coverage of mappings for SNOMED-CT was the following. The Chest+COVID dataset contains a total of 299 concepts. Of these, 245 (81.9%) had been assigned a UMLS CUI, and of these 245, 177 (72.24%) had a mapping from UMLS CUI to SNOMED-CT available at UMLS terminology service. The remaining 68 had no mapping available to SNOMED-CT (27.75%). Of these 68, 54 could be mapped ad hoc to one SNOMED-CT US concept, eight had to be mapped to a more general concept in SNOMED-CT, four required postcoordination to express their semantics, and two did not have a candidate nor could they be expressed using postcoordination. Those requiring postcoordination for the anatomical qualifier were implemented with the pattern described in the following sections. Non-anatomical qualifiers for these two concepts were disregarded. Beyond SNOMED-CT, concepts were mapped to more domain-specific terminologies when applicable. Specifically, 18 concepts were mapped to ICD-10, 49 were mapped to HPO, 35 were mapped to ICPC-2, 63 were mapped to MDRSPA, 18 were mapped to MEDLINEPLUS, 69 terms were mapped to FMA. The remaining terminologies were included but had no mapping from Chest+COVID (ATC, LOINC, NCI PI-RADS, caDSR). The reason is that BIMCV also includes images captured by other methods and body parts, such as those related to prostate cancer.

After terminology mapping (2), it is possible to identify equivalent terms to implement phenotyping queries. However, it is not possible to reason over the mappings to, for example, deal with *is-a* (subsumption) relationships among concepts querying for a given concept and letting the DB to retrieve all the children of that concept. To tackle this issue, Fig. 4 shows in (3) how the resulting LKB formalizes all the relationships among the terminologies supported. First, the SNOMED-CT lite RDF(S) (see white ellipses in Fig. 4) version was generated to have a complete and sound organization of clinical concepts with a formal logic model that allows for reasoning in a triple store DB. Second, all terminology mappings produced in (2) were processed. An *owl:equivalentClass* axiom was produced for each mapping, relating equivalent concepts among the terminologies supported. Fig. 4 represents HPO concepts with gray hexagons, FMA concepts with yellow hexagons, and UMLS concepts with blue rectangles. For each report, a relationship *lkb:referencesLocationConcept* was produced, linking each report to the anatomical locations it references. The development of this relationship is a decision aiming for interoperability among triple stores DBs. Although the SNOMED-CT concept model allows qualifying anatomical locators using postcoordinated expressions, that would constrain the set of triple stores that could be used to process these expressions. Hence, we opted to model references to anatomical locations with an explicit relationship. The phenotyping queries developed in SPARQL code are available in Appendix C. Fig. 4 shows an excerpt of the LKB to explain how the DB reasons over the LKB to answer Query 1 and Query 2 from Table 7.

The execution of Query 1 *"retrieve all radiologic reports that contain a Pleural effusion in the left lung"* in Fig. 4 is depicted by highlighting in green the properties and OWL axioms analyzed. Query 1 uses the HPO concept *HP:0002202 (Pleural effusion)*, which is equivalent to the concept *C0032227 (Pleural effusion)* in UMLS. By analyzing this equivalence relationship, the DB will determine that the concept *HP:0002202 (Pleural effusion)* is indirectly referenced by the report with identifier *4991845* through the *referencesConcept* relationship of concept UMLS: C0032227. With regards to the location of the finding (i.e., left lung), the DB will infer that the concept in SNOMED-CT *44029006 (Left lung structure)* is equivalent to the FMA concept *7310 (Left Lung)* (referenced in the query) by transitively analyzing the equivalences of UMLS: C0225730. In addition, the DB will determine by subsumption that SNOMED-CT 41224006 is also an FMA:7310 (concept referenced in the query). Since *SNOMED-CT 41224006* is referenced by the report through the *referencesLocationConcept* property in UMLS: C0225749 (Structure of posterior basal segment of the left lobe of the lung), the report will be eligible to be returned as a query result. In this way, both the finding and the anatomical location referenced in the query will be considered, and the report identifier 4765778 will be returned.

Fig. 4 shows the execution of Query 2: *"Retrieve those images where it is observed a radiologic infiltrate in a bronchopulmonary segment structure"*. The figure depicts the properties and OWL axioms analyzed with blue arrows to determine the query results. To answer this query, the triple store DB reasons over the LKB determining that the concept in the HPO *HP:0002113 (radiologic infiltrate)* is equivalent to UMLS concept *C1265599* which, in turn, is referenced by the radiology report with identifier *4765778*. Additionally, it determines the valid locations for the concept *radiologic infiltrate* by: (a) analyzing that the SNOMED-CT concept *46148009 (Structure of posterior basal segment of the left lower lobe of the lung)* is a *subClassOf SNOMED-CT 72674008 Bronchopulmonary segment structure*; and, (b) analyzing that *46148009 is equivalentTo C0225749* which is directly referenced by the report with identifier *4765778*.

Fig. 5 depicts the architecture used to execute queries. Stage I represents the NLP processing and ontology development, resulting in the LKB stored as a graph in the triple store. Stage II represents executing a query over the triple store to retrieve a report. Specifically, the clinical data scientist on top of Fig. 5 types a query using the SPARQL language to retrieve the radiologic reports referencing the disorder *"pleural effusion"* and the anatomical location *"left lung"* (see WHERE section of the query). As explained previously, the triple store (Graph database) executes the query transitively, analyzing all the relationships in the LKB. The report that complies with the phenotyping criteria is then returned to the user.

The LKB developed aims to go beyond the design of an ad-hoc knowledge graph by maximizing its interoperability and ontology reuse [39]. This was achieved by not only committing to biomedical ontologies but to general metadata ontologies. Specifically, we reused concepts from schema.org [40], the Dublin Core [41] and the Minimal Service Model [42]. From schema.org MedicalProcedure was subclassed to specify the Report concept, MedicalStudy to identify the specific study responsible for eliciting this report, and Patient to model patients metadata such as pseudonym, age, and gender. ImageObject to model DICOM image metadata. MSM was used to link the ontological definition of clinical and report-related concepts (study, patient etc.) to the syntactic layer of the Web service, specifying where the endpoint is and the message produced by the Web service. Dublin Core was used for metadata location such as provenance, data set creator, etc. The design based on ontology reuse and Linked Data facilitates the compliance with Findability, Accessibility, Interoperability, and Reusability (FAIR) principles [43] and Open Science [44] by allowing the publication, discovery, and exploration of the LKB by computers.

5. Discussion

5.1. NLP and transformer-based models

There are several studies making use of transformers for NLP on clinical text. EHR text may lead to different performances than medical literature or general language when using transformers. In [45], the authors evaluated word embeddings trained from different types of corpora (EHR, medical literature, Wikipedia, and news). In their results, clinical notes from the EHR and medical literature captured the semantics of medical terms in the way closest to human judgment. In [46], the authors proposed a framework as a preliminary step to understand textual spatial semantics in chest X-ray reports. In their framework, common radiological entities tied to spatial relations are encoded through four spatial roles: trajector, landmark, diagnosis, and hedge, all identified in relation to a spatial preposition (or spatial indicator). Their study used bidirectional long-short term memory (Bi-LSTM) as a base model and compared it with two models using transformers (BERT and XLNet). The best-performing models were both XLNet and BERT, pre-trained with the medical dataset MIMIC-III. A

Table 7
List of phenotyping queries implemented with their description, terminologies used and type of reasoning to find query results. See C for SPARQL implementation.

Query id	Text definition	Reasoning
Query 1	Retrieve all radiologic images that contain a Pleural effusion in the left lung	Subsumption over SNOMED-CT, equivalence among HPO, UMLS, SNOMED-CT and FMA
Query 2	Retrieve those images where it is observed a radiologic infiltrate in a bronchopulmonary segment structure	Subsumption over SNOMED-CT, equivalence among HPO, UMLS and SNOMED-CT
Query 3	Retrieve patients with widened mediastinum	Subsumption over SNOMED-CT, equivalence between UMLS and SNOMED-CT
Query 4	Retrieve all reports with any radiology finding observed in the mediastinal area	Subsumption over SNOMED-CT, equivalence between UMLS and SNOMED-CT
Query 5	Retrieves all images that have some type of catheter	Subsumption over SNOMED-CT, equivalence between UMLS and SNOMED-CT
Query 6	Retrieves all images where fractures are observed in the shoulder region	Subsumption over SNOMED-CT, equivalence among FMA, UMLS and SNOMED-CT
Query 7	recover images with multiple nodules and pseudonodules	Subsumption over SNOMED-CT, equivalence among HPO, UMLS and SNOMED-CT
Query 8	Recover images with a finding of atelectasis	Subsumption over SNOMED-CT, equivalence among ICD-10, UMLS, SNOMED-CT

subdomain that has received lots of attention is the processing of free text medication orders. In [47], GloVe word vectors trained from MIMIC III were used, and four models for supervised classification were tested (Multinomial Naive Bayes, Decision Trees, Support Vector Machines, and CRFs). In a second step, the medication items (dosage, mode, frequency etc.) were extracted. Multi-label CRF was selected as the optimal method for medication information extraction. Their work did not perform terminology matching for phenotyping.

Several works have worked in NLP on pathology reports to identify location and morphology of tumors. Oleynic et al. [48] used Support Vector Machines to classify pathology reports into ICD-O. Mitchel et al. [49] proposed a 3-stage method for processing pathology reports. In the first stage, they created a pathology language model using BioBERT and a subset of MIMIC III concerning discharge summaries. In the second stage, they used SQuAD and BioASQ datasets to train a question answering model that could retrieve the organs containing a tumor and the kind of tumor. In the third stage, they trained 2 additional models to predict the ICD-O-3 codes that corresponded to the location and type of tumor retrieved in the second stage. Our methodology has not used question answering models such as SQuAD. A relevant future work is the comparison of the performance of these methods with ours to map to UMLS CUIs. Also, in the pathology domain, Rios et al. [50] used a neural multi-task training with hierarchical regularization to process pathology reports to integrate concept embeddings into the ICD-O-3 hierarchy. This technique could be used in our pipeline to automatically process concepts without SNOMED-CT mapping in UMLS and link them to the correct concept in the SNOMED-CT hierarchy.

Coutinho and Martins [11] used BERT-based models to map causes of death from death certificates in Portuguese into their corresponding ICD-10 codes. Their work used a BERT model trained on the Brazilian Wikipedia data and the brWac dataset. Adjustment of parameters to clinical text was made with self-supervised pre-training tasks. The authors used a novel pre-training procedure that incorporates in-domain knowledge, and also a fine-tuning method to address the class imbalance issue. Experimental results show that, in this particular clinical task that requires the processing of relatively short documents, Transformer-based models can achieve very competitive results. This work presents quite a few similarities with the part of our work that performs the labeling from the free text of the EHRs. There are also recent works based on different machine learning methods to obtain classifiers. López-Úbeda and colleagues [51] develop a text classification system based on NLP in order to automatically assign one protocol to each radiological request form prescribed by reference physicians. For this task they used two training Spanish corpora. Olthof and colleagues [52] classify radiology reports in orthopedic trauma for the presence of injuries. They used a dataset of Dutch radiology reports of injured extremities and a dataset of chest radiographs.

Another feature of our study is using transformers on a Spanish clinical corpus. Most of the studies mentioned were done in English on similar corpora (e.g. MIMIC III, Medical literature, UW Dataset). Our study demonstrates that transformer-based methods can be applied to other languages, such as Spanish, while maintaining a good performance.

5.2. Linked knowledge base design

Concerning the design of the LKB, a first difference in the treatment of relationships between concepts can be found between our study and Datta et al. [46]. Datta and colleagues focused on identifying spatial relationships, but terminology linkage still needed to be performed. In our study, we did not perform spatial relationship analysis at the NLP stage. However, we dealt with anatomical locations by mapping them to FMA and SNOMED-CT, thus allowing for reasoning over partonomies. For example, if a phenotyping query asks for the radiology reports that mentioned any fracture in the region of the shoulder; reports mentioning fractures in the acromioclavicular joint, rotator cuff, supraspisnous humeral head, humeral neck and so on, will be retrieved. The reason is that these concepts are related by is-a (rdfs:SubClassOf) relationships with the *shoulder region structure (16982005)*. Another design decision that differs from other works is the specification of the semantics regarding clinical findings and anatomical locations using *lkb:referencesConcept* and *lkb:referencesLocationConcept* relationships, respectively. In many cases, only the former may be needed since the clinical finding hierarchy (404684003) in SNOMED-CT contains precoordinated terms that directly relate the finding and the anatomical location (e.g., *445249002 | Multiple nodules of lung (disorder)*). However, there are cases where a precoordinated term specifying both the clinical finding and the anatomical location is unavailable. A possible approach to this may be using the postcoordination mechanism provided by SNOMED-CT. However, this may hamper interoperability. While some reasoners perform well in classifying postcoordinated concepts [53], not all of them support the semantics derived from postcoordinated constructs. In our design, we adopted a more flexible approach where the anatomical location in the finding can be expressed by: (a) directly referencing a precoordinated concept with anatomical location (*lkb:referencesConcept*); or (b) using the relationship *lkb:referencesLocationConcept* to specify the anatomical location. This improves interoperability across triple stores and facilitates mapping to domain-specific terminologies such as the FMA following Linked Data principles. Other studies have explored using

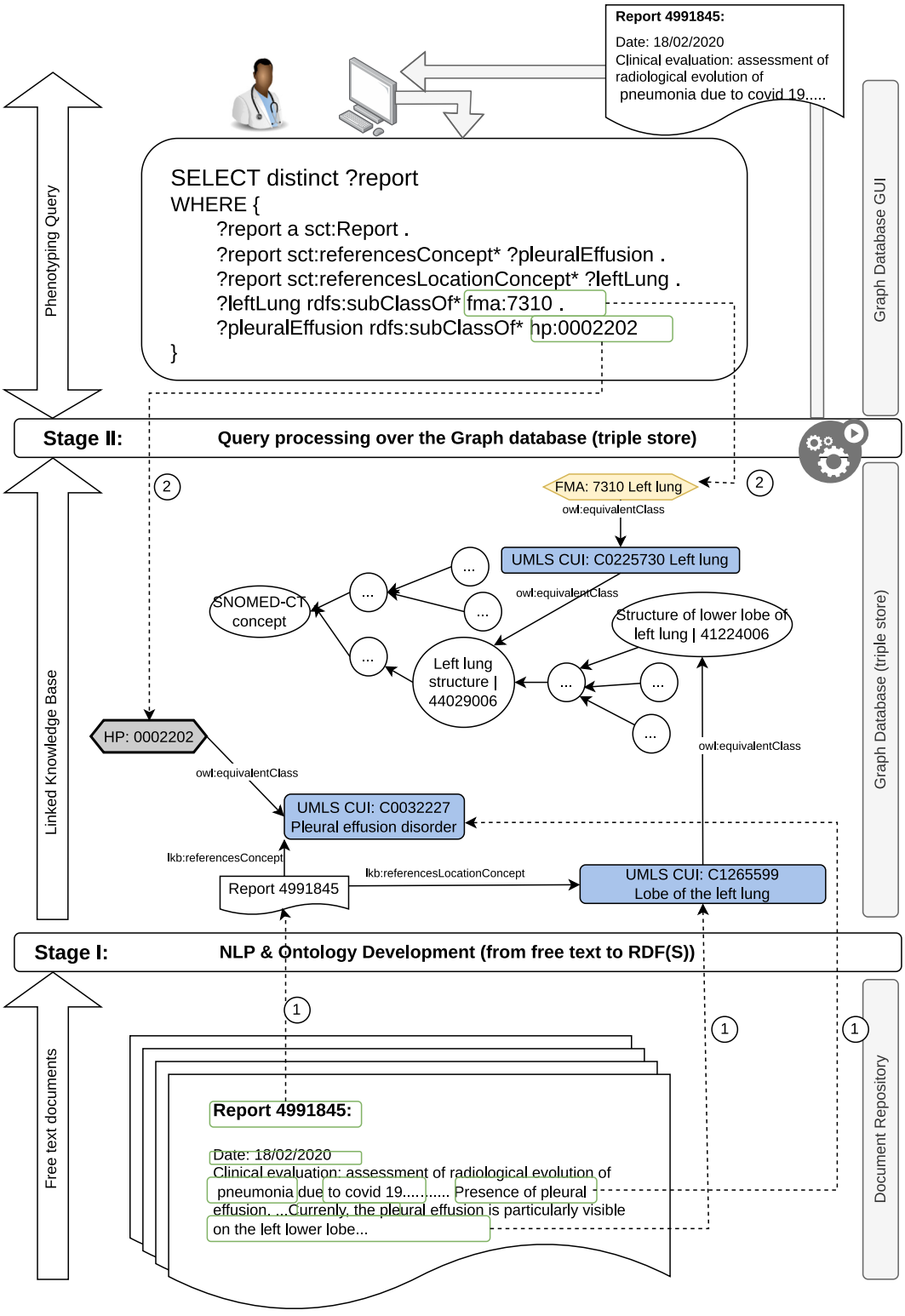

Fig. 5. Query architecture on the graph Data Base.

Linked Data to increase the interoperability of data reuse frameworks such as OHDSI [54]. However, the use of Linked Data was focused on mappings from the internal terminologies of the framework to more interoperable ones such as UMLS and HPO. Conversely, our work focused on direct ontology learning from free text, focusing on the accuracy and correct taxonomic classification of the identified entities. Also, Banda et al. [55] used Machine Learning to learn phenotypes from imperfect labeled data, but it was done by reading over the OHDSI CDM v5 structured model rather than free text. They constrained the XPRESS model [56] to operate only over standard terms rather than

free text [55]. The work closest to ours in terms of mapping text to an ontology is the study of Hammami et al. [34], which used a rule-based system to map pathology reports in Italian to the International Classification of Diseases for Oncology (ICD-O-M). Hammami and colleagues dealt with classifying concepts into the ICD-O-M using rules to prioritize morphology descriptions. However, the ontology produced was not formalized into a logic model allowing the execution of phenotyping queries.

5.3. Limitations

Our method allows us to process free text notes directly and define phenotypes to filter radiology reports using various terminologies. However, phenotypes are as diverse as the natural language contained in reports, and our methods have several limitations regarding the expressivity they can achieve.

First, Deep Phenotyping requires not only the specification of clinical concepts but also the interpretation of temporal constraints and demographic data in free text. Our methodology does not deal with these features. That challenge would require the development of temporal NLP models in addition to the ones presented [57], which remain as future work.

Second, the methods presented are not a complete clinical research platform such as i2b2 or OHDSI that can define more precise phenotypes (e.g., with temporal constraints) at the expense of intensive ETL and data curation efforts. Instead, the methodology presented is intended to query large volumes of clinical reports for assigning them to clinical studies with minimum ETL efforts. Once radiology reports have been retrieved by running phenotyping queries with the patterns presented, each study can decide to normalize clinical report data further into clinical research platforms for further analysis.

Third, our methodology alleviates most of the effort in ETL tasks. However, the functionality of our method is limited to answering queries for detecting clinical reports with the patterns specified. That is, the phenotyping criteria that can be processed are limited by the completeness of the LKB in expressing the contents of free text reports. Dealing with more expressive criteria would require the NLP pipeline to be enhanced to detect more complex relationships and then translate those to the LKB as logic constructs. Nevertheless, the management of clinical reports and the phenotyping queries supported suffice the needs of BIMCV regarding indexing and fast processing of reports. This is aligned with the findings of Sholle et al. [58], who reported that only 15% of phenotyping queries required more than three conditions or custom temporal constraints. Hence, we believe our method can be reliably used directly in various use cases, such as image and pathology banks, as long as representative corpora are developed.

Finally, including more domain-specific terminologies, such as RadLex [59], can enhance interoperability and cross-terminological phenotyping queries. However, in the context of public healthcare in Spain, RadLex is not used. Thus, its adoption remains as future work.

6. Conclusions

EHRs contain large amounts of free text. This study presented a methodology that leverages the use of transformer-based NLP methods with Linked Data Technologies to enable Deep Phenotyping on clinical reports. Our phenotyping system was evaluated on the Spanish Chest+COVID dataset showing state-of-the-art results for the NLP task and effectively supporting expressive phenotyping queries. The performance of the multi-label classification module was very satisfactory to reinforce Deep Phenotyping, achieving a micro averaging of 0.950, a macro averaging of 0.691, and a weighted average of 0.949. To answer Deep Phenotyping queries, 12 terminologies were mapped and expressed in compliance with Linked Data principles as a LKB that effectively processed equivalence and subsumption relationships. Subsequent work will involve incorporating temporal evolution into the NLP module and extending our NLP-LKB system to other domains. Additionally, it would be interesting to have a gold standard to carry out a quantitative and qualitative analysis of the benefits of the complete system. The results could show, for example, the relationship between the errors in the labeling phase from the medical reports in terms of the CUIs labels and the results of the phenotyping queries. The study also notes the rising interest in cross-lingual knowledge transfer strategies for low-resource languages [60–62] and in Biomedical or Medical entity linking [63–65], with an exploration of these strategies identified as a direction for future work.

Finally, Chest+COVID is a publicly open dataset, and our intention is to make the corpus publicly available to adhere to FAIR principles [43]. This step aligns with EU recommendations on Open Science [44] by indexing our results in the European Open Science Cloud and the Linking Open Data Cloud.

CRediT authorship contribution statement

Lluís-F. Hurtado: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Luis Marco-Ruiz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Encarna Segarra:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Maria Jose Castro-Bleda:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Aurelia Bustos-Moreno:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria de la Iglesia-Vayá:** Writing – original draft, Resources, Project administration, Data curation, Conceptualization. **Juan Francisco Vallalta-Rueda:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare not having any conflict of interest.

Acknowledgments

This work has been carried out during phase I of the Big Data Personalized Medicine project, file 18/CPP/1. The Big Data Personalized Medicine project has been co-financed by the European Regional Development Fund (FEDER) in 85% for the €3,833,774 budgeted for the Canary Islands Health Service, and 50% for the €2,000,000 budgeted for the Generalitat Valenciana, through a grant granted by the Ministry of Science and Innovation of €4,258,707.90. It is also funded by the Ministerio de Ciencia e Innovación, Spain and European Union under project BEWORD PID2021-126061OB-C41, and by the Generalitat Valenciana, Spain under projects CIPROM/2021/023 and PROMETEO/2020/024.

Appendix A. Experimental results

Results for each label on the Test set, in terms of true positives (tp), false positives (fp), false negatives (fn), Precision (P), Recall (R), F1-score, and Support (S).

Label	tp	fp	fn	P	R	F1	S
normal	168	5	8	0.971	0.955	0.963	176
unchanged	125	3	4	0.977	0.969	0.973	129
exclude	95	10	14	0.905	0.872	0.888	109
pneumonia	78	7	2	0.918	0.975	0.945	80
interstitial pattern	71	1	1	0.986	0.986	0.986	72
pleural effusion	67	1	1	0.985	0.985	0.985	68
alveolar pattern	60	1	1	0.984	0.984	0.984	61
infiltrates	44	3	2	0.936	0.957	0.946	46
costophrenic angle blunting	44	2	0	0.957	1.000	0.978	44
laminar atelectasis	41	0	0	1.000	1.000	1.000	41
atelectasis	36	0	1	1.000	0.973	0.986	37
chronic changes	34	0	0	1.000	1.000	1.000	34
callus rib fracture	30	0	1	1.000	0.968	0.984	31
increased density	25	3	6	0.893	0.806	0.847	31
cardiomegaly	27	0	1	1.000	0.964	0.982	28
pseudonodule	26	0	1	1.000	0.963	0.981	27
vertebral degenerative changes	25	0	2	1.000	0.926	0.962	27
nodule	23	2	3	0.920	0.885	0.902	26
scoliosis	26	0	0	1.000	1.000	1.000	26
COPD signs	22	1	1	0.957	0.957	0.957	23
calcified granuloma	23	1	0	0.958	1.000	0.979	23
heart insufficiency	21	0	1	1.000	0.955	0.977	22
volume loss	22	1	0	0.957	1.000	0.978	22
consolidation	19	2	2	0.905	0.905	0.905	21
vascular hilar enlargement	19	1	1	0.950	0.950	0.950	20
air trapping	18	0	1	1.000	0.947	0.973	19
bronchovascular markings	18	2	0	0.900	1.000	0.947	18
suboptimal study	15	1	3	0.938	0.833	0.882	18
apical pleural thickening	17	2	0	0.895	1.000	0.944	17
fibrotic band	16	0	1	1.000	0.941	0.970	17
kyphosis	16	2	0	0.889	1.000	0.941	16
bronchiectasis	15	0	0	1.000	1.000	1.000	15
COVID 19	14	1	0	0.933	1.000	0.966	14
aortic elongation	14	0	0	1.000	1.000	1.000	14
nipple shadow	14	0	0	1.000	1.000	1.000	14
pacemaker	13	0	0	1.000	1.000	1.000	13
suture material	12	1	1	0.923	0.923	0.923	13
vertebral anterior compression	13	2	0	0.867	1.000	0.929	13
hemidiaphragm elevation	11	2	1	0.846	0.917	0.880	12
NSG tube	10	1	1	0.909	0.909	0.909	11
aortic atheromatosis	8	0	2	1.000	0.800	0.889	10
calcified densities	9	0	1	1.000	0.900	0.947	10
emphysema	10	1	0	0.909	1.000	0.952	10
endotracheal tube	10	0	0	1.000	1.000	1.000	10
hilar congestion	10	0	0	1.000	1.000	1.000	10
hilar enlargement	8	0	2	1.000	0.800	0.889	10
granuloma	9	1	0	0.900	1.000	0.947	9
flattened diaphragm	8	0	0	1.000	1.000	1.000	8
goiter	7	0	1	1.000	0.875	0.933	8
ground glass pattern	8	0	0	1.000	1.000	1.000	8
osteosynthesis material	8	0	0	1.000	1.000	1.000	8
superior mediastinal enlargement	8	0	0	1.000	1.000	1.000	8

adenopathy	7	0	0	1.000	1.000	1.000	7
central venous catheter via jugular vein	7	0	0	1.000	1.000	1.000	7
mediastinic lipomatosis	6	0	1	1.000	0.857	0.923	7
metal	7	1	0	0.875	1.000	0.933	7
pneumothorax	7	0	0	1.000	1.000	1.000	7
rib fracture	7	0	0	1.000	1.000	1.000	7
vertebral compression	6	0	1	1.000	0.857	0.923	7
dai	5	1	1	0.833	0.833	0.833	6
diaphragmatic eventration	5	0	1	1.000	0.833	0.909	6
dual chamber device	6	1	0	0.857	1.000	0.923	6
lobar atelectasis	6	0	0	1.000	1.000	1.000	6
multiple nodules	4	1	2	0.800	0.667	0.727	6
surgery breast	6	0	0	1.000	1.000	1.000	6
tracheal shift	6	0	0	1.000	1.000	1.000	6
COVID 19 uncertain	3	0	2	1.000	0.600	0.750	5
aortic button enlargement	4	0	1	1.000	0.800	0.889	5
bullas	5	0	0	1.000	1.000	1.000	5
hiatal hernia	5	1	0	0.833	1.000	0.909	5
mediastinal enlargement	4	0	1	1.000	0.800	0.889	5
reservoir central venous catheter	5	0	0	1.000	1.000	1.000	5
sternotomy	4	0	1	1.000	0.800	0.889	5
supra aortic elongation	5	1	0	0.833	1.000	0.909	5
tuberculosis sequelae	5	1	0	0.833	1.000	0.909	5
vascular redistribution	5	0	0	1.000	1.000	1.000	5
atypical pneumonia	4	0	0	1.000	1.000	1.000	4
bone metastasis	2	0	2	1.000	0.500	0.667	4
calcified pleural thickening	4	0	0	1.000	1.000	1.000	4
cavitation	4	0	0	1.000	1.000	1.000	4
clavicle fracture	4	0	0	1.000	1.000	1.000	4
humeral fracture	3	0	1	1.000	0.750	0.857	4
lung metastasis	3	0	1	1.000	0.750	0.857	4
lung vascular paucity	2	0	2	1.000	0.500	0.667	4
mastectomy	4	0	0	1.000	1.000	1.000	4
osteoporosis	4	0	0	1.000	1.000	1.000	4
sclerotic bone lesion	4	1	0	0.800	1.000	0.889	4
single chamber device	4	0	0	1.000	1.000	1.000	4
vertebral fracture	3	0	1	1.000	0.750	0.857	4
ascendent aortic elongation	2	1	1	0.667	0.667	0.667	3
axial hyperostosis	2	0	1	1.000	0.667	0.800	3
central venous catheter	3	0	0	1.000	1.000	1.000	3
chest drain tube	3	0	0	1.000	1.000	1.000	3
end on vessel	2	0	1	1.000	0.667	0.800	3
loculated pleural effusion	2	0	1	1.000	0.667	0.800	3
miliary opacities	3	1	0	0.750	1.000	0.857	3
pleural thickening	3	1	0	0.750	1.000	0.857	3
pulmonary edema	3	0	0	1.000	1.000	1.000	3
pulmonary mass	2	1	1	0.667	0.667	0.667	3
surgery lung	1	0	2	1.000	0.333	0.500	3
surgery neck	3	0	0	1.000	1.000	1.000	3
abnormal foreign body	2	0	0	1.000	1.000	1.000	2
air bronchogram	2	0	0	1.000	1.000	1.000	2
azygoesophageal recess shift	2	1	0	0.667	1.000	0.800	2
blastic bone lesion	1	0	1	1.000	0.500	0.667	2
costochondral junction hypertrophy	2	0	0	1.000	1.000	1.000	2
hypoexpansion	2	0	0	1.000	1.000	1.000	2
lytic bone lesion	2	0	0	1.000	1.000	1.000	2
major fissure thickening	2	0	0	1.000	1.000	1.000	2
minor fissure thickening	2	0	0	1.000	1.000	1.000	2
osteopenia	2	0	0	1.000	1.000	1.000	2
subacromial space narrowing	2	0	0	1.000	1.000	1.000	2

subcutaneous emphysema	1	0	1	1.000	0.500	0.667	2
tuberculosis	1	0	1	1.000	0.500	0.667	2
air fluid level	1	0	0	1.000	1.000	1.000	1
artificial aortic heart valve	1	0	0	1.000	1.000	1.000	1
artificial heart valve	1	0	0	1.000	1.000	1.000	1
artificial mitral heart valve	1	0	0	1.000	1.000	1.000	1
atelectasis basal	1	0	0	1.000	1.000	1.000	1
azygos lobe	1	0	0	1.000	1.000	1.000	1
calcified adenopathy	1	1	0	0.500	1.000	0.667	1
calcified fibroadenoma	1	0	0	1.000	1.000	1.000	1
central venous catheter via subclavian vein	1	0	0	1.000	1.000	1.000	1
central venous catheter via umbilical vein	1	0	0	1.000	1.000	1.000	1
descendent aortic elongation	1	0	0	1.000	1.000	1.000	1
endoprosthesis	1	0	0	1.000	1.000	1.000	1
fissure thickening	0	0	1	0.000	0.000	0.000	1
gastrostomy tube	1	0	0	1.000	1.000	1.000	1
gynecomastia	1	0	0	1.000	1.000	1.000	1
heart valve calcified	1	0	0	1.000	1.000	1.000	1
humeral prosthesis	1	0	0	1.000	1.000	1.000	1
hyperinflated lung	1	1	0	0.500	1.000	0.667	1
mammary prosthesis	0	0	1	0.000	0.000	0.000	1
mediastinal mass	0	0	1	0.000	0.000	0.000	1
pectum carinatum	1	0	0	1.000	1.000	1.000	1
pericardial effusion	1	0	0	1.000	1.000	1.000	1
pleural plaques	1	0	0	1.000	1.000	1.000	1
pneumomediastinum	0	0	1	0.000	0.000	0.000	1
pneumoperitoneo	1	0	0	1.000	1.000	1.000	1
post radiotherapy changes	1	0	0	1.000	1.000	1.000	1
prosthesis	1	0	0	1.000	1.000	1.000	1
pulmonary fibrosis	1	0	0	1.000	1.000	1.000	1
pulmonary hypertension	1	0	0	1.000	1.000	1.000	1
respiratory distress	0	0	1	0.000	0.000	0.000	1
reticular interstitial pattern	1	1	0	0.500	1.000	0.667	1
reticulonodular interstitial pattern	0	0	1	0.000	0.000	0.000	1
segmental atelectasis	1	0	0	1.000	1.000	1.000	1
soft tissue mass	1	0	0	1.000	1.000	1.000	1
surgery	0	1	1	0.000	0.000	0.000	1
thoracic cage deformation	1	0	0	1.000	1.000	1.000	1
ventriculoperitoneal drain tube	1	0	0	1.000	1.000	1.000	1
non axial articular degenerative changes	0	2	0	0.000	0.000	0.000	0
Chilaiditi sign	0	1	0	0.000	0.000	0.000	0

Appendix B. Ad hoc mappings

UMLS CUI	UMLS definition	SNOMED-CT mapping	Comment
C3544344	Ground glass opacity on chest X-ray	1217294009 | Ground glass lung opacity (finding)	
C2073636	x-ray of chest: pneumomediastinum-finding	16838000 | Mediastinal emphysema (disorder)	
C2203586	x-ray: calcifications	129748009 | Radiographic calcification finding (finding)	generalized to parent concept
C0546312	Hyperexpansion of lung	249674001 | Chest over-expanded (finding)	
C0740844	AIR FLUID LEVEL (finding)	NO CANDIDATE	
C2073538	x-ray of chest: interstitial infiltrate of lung	409609008 | Radiologic infiltrate of lung (disorder)	
C2026143	central venous line with subcutaneous reservoir	52124006 | Central venous catheter, device (physical object)	
C1399223	presence; artificial heart valve	161677002 | History of artificial heart valve (situation)	
C4477098	Pulmonary venous hypertension	5499009 | Pulmonary hypertensive venous disease (disorder)	generalized to parent concept
C2073565	x-ray of chest: lungs pneumothorax	36118008 | Pneumothorax (disorder)	
C2828075	Suboptimal Image Reason	<< 366056003 |Finding of quality of visual image (finding)| : << 363713009 |Has interpretation (attribute)| = << 423437008 |Insufficient (qualifier value)	postcoordination necesary
C1851119	Aortic arch dilatation	373134008 | Aneurysm of aortic root (finding)	
C0239041	CHEST XRAY PULMONARY VASCULAR REDISTRIBUTION	15232001 | Increased vascular markings of lung (finding)	
C1096249	Calcification of the aorta	250978003 | Aortic valve calcification (disorder)	
C2203581	x-ray: blastic bone metastasis	712849003 | Primary malignant neoplasm of prostate metastatic to bone (disorder)	generalized to parent concept since no qualifier for "blastic" was found
C0742855	COSTOPHRENIC ANGLE OBLITERATED	<< 119271006Obliteration (procedure)| : << 363704007Procedure site (attribute)| = << 46297007Structure of costophrenic angle (body structure)	postcoordination necesary
C3889085	Ascending aortic dilatation	425963007 | Aneurysm of ascending aorta (disorder)	
C2073672	x-ray of chest: reticular-nodular interstitial infiltrate	409609008 | Radiologic infiltrate of lung (disorder)	
C4273001	cxr mediastinum widening superior	363646005 | Widened mediastinum (finding)	
C2115817	kyphosis	414564002 | Kyphosis deformity of spine (disorder)	
C4049711	Lepidic Predominant Adenocarcinoma	128660005 | Bronchiolo-alveolar carcinoma, mucinous (morphologic abnormality)	
C0239019	CHEST XRAY KERLEY A,B, OR C LINES	NO CANDIDATE	
C1400000	vertebra; hyperostosis	129138001 | Disorder of thoracic spine (disorder)	generalized to parent concept
C0917968	Mammary Prostheses, unspecified whether internal or external	2282003 | Prosthetic breast implant (physical object)	
C2073707	x-ray of chest: unilateral elevation of diaphragm	88936002 | Elevated diaphragm (disorder)	
C0869748	heart - aortic valve	34202007 | Aortic valve structure (body structure)	
C2021206	x-ray of chest: mediastinal widening	363646005 | Widened mediastinum (finding)	
C4476542	obsolete Dilatation of the descending aorta	449184004 | Dilatation of descending aorta (disorder)	
C3178780	Chilaiditi Anomaly	14911005 | Subphrenic interposition syndrome (disorder)	
C0457196	Soft tissue mass of thorax	444905003 | Mass of soft tissue (finding)	
C0024881	Mastectomy	69031006 | Excision of breast tissue (procedure)	
C2073504	x-ray of chest: flattened diaphragm	106062002 | Diaphragmatic finding (finding)	
C0745058	HUMERUS PROSTHESIS	705881009 | Elbow humerus prosthesis (physical object)	
C2073583	x-ray of chest: miliary infiltrate of lung	409609008 | Radiologic infiltrate of lung (disorder)	
C0869752	heart - mitral valve	91134007 | Mitral valve structure (body structure)	
C4290224	arthrosis or osteoarthritis of spine	8847002 | Spondylosis (disorder)	
C1332240	Abnormal alveolar pattern	409609008 | Radiologic infiltrate of lung (disorder)	generalized to parent concept
C4315325	Sclerotic bones	37748009 | Bony sclerosis (morphologic abnormality)	
C0972395	Automatic Implantable Cardioverter-Defibrillators	118381000 | Implantable cardioverter leads, device (physical object)	
C2073448	x-ray of chest: calcification of heart valve	373136005 | Heart valve calcification (finding)	
C2073563	x-ray of chest: lungs multiple pulmonary nodules	445249002 | Multiple nodules of lung (finding)	
C0865843	Atrophic fibrosis of lung chronic or unspecified	13274008 | Atrophic fibrosis of lung (disorder)	

C1698506	Pulmonary hilar enlargement	<< 274710003 |Lung field abnormal (finding)| : << 363698007 |Finding site (attribute)| = << 46750007 |Structure of hilum of lung (body structure)|, << 116676008 |Associated morphology (attribute)| = << 442021009 |Enlargement (morphologic abnormality)	postcoordination necesary
C2073485	x-ray of chest: cyst of lung	275504005 | Cyst of lung (disorder)	
C2073625	x-ray of chest: pleural effusion	60046008 | Pleural effusion (disorder)	
C0742362	CHEST X RAY ABNORMAL CHRONIC	442147002 |Imaging of thorax abnormal (finding)| : << 263502005 |Clinical course (attribute)| = << 90734009 |Chronic (qualifier value)|, << 363698007 |Finding site (attribute)| = << 51185008 |Thoracic structure (body structure)	postcoordination necesary
C0746053	LUNG BASE ATELECTASIS	46621007 | Atelectasis (disorder)	
C2073518	x-ray of chest: increased bronchovascular markings	15232001 | Increased vascular markings of lung (finding)	
C2072932	Dilated pulmonary arteries	251047005 | Dilatation of pulmonary artery (disorder)	
C4538889	Thoracic deformation	448186009 | Deformity of thoracic structure (disorder)	
C1333298	Diffuse Lipomatosis	402693001 | Lipomatosis (disorder)	
C0734296	Apical zone of lung	245515008 | Structure of apical segment of upper lobe of lung (body structure)	
C0934569	Hemithoracic structure	422614002 | Structure of half of thorax lateral to midsagittal plane (body structure)	
C0504099	Right costodiaphragmatic recess	6731002 | Structure of pleural recess (body structure)	generalized to parent concept
C0504100	Left costodiaphragmatic recess	6731002 | Structure of pleural recess (body structure)	generalized to parent concept
C1522619	Esophageal	32849002 | Esophageal structure (body structure)	
C1522601	Cardiac - anatomy qualifier	119202000 | Heart part (body structure)	
C0934260	Suprahilar part of mediastinal pleura	27416003 | Mediastinal pleura structure (body structure)	generalized to parent concept, qualifier for " suprahiliar" was not found
C0929434	Middle lung field	281393007 | Structure of middle zone of lung (body structure)	
C0205150	Hilar	46750007 | Structure of hilum of lung (body structure)	
C4253583	Oblique fissure of lung (line)	278983006 | Structure of fissure of lung (body structure)	
C1261193	Inferior mediastinum	1187376008 | Structure of inferior mediastinum (body structure)	
C0929165	Bronchopulmonary subsegment	72674008 | Bronchopulmonary segment structure (body structure)	
C4323265	Posterior third of body of rib	263362007 | Structure of posterior third of shaft of rib (body structure)	
C0920882	Region of neck (body structure)	298378000 | Finding of neck region (finding)	
C4323264	Anterior third of body of rib	728543004 | Entire anterior third of shaft of rib (body structure)	
C0929227	Upper lung field	281392002 | Structure of upper zone of lung (body structure)	
C1522318	Coronary	41801008 | Coronary artery structure (body structure)	

Color codes:
Postcoordination is necesary
No candidate could be found
The concept had to be generalized to a concept with a more general meaning

Appendix C. Phenotyping queries

Query 1-Retrieve all radiologic reports that contain a Pleural effusion in the left lung

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
PREFIX fma: <http://purl.org/sig/ont/fma/>
PREFIX hp: <http://purl.obolibrary.org/obo/hp#>
SELECT distinct ?report
WHERE {
  ?report a sct:Report .
  ?report sct:referencesConcept* ?pleuralEffusion .
  ?report sct:referencesLocationConcept* ?leftLung .
  ?leftLung rdfs:subClassOf* fma:7310 .
  ?pleuralEffusion rdfs:subClassOf* hp:0002113
}
```

Query 2-Retrieve reports mentioning a pattern of infiltration in a bronchopulmonary segment structure.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
```

```
PREFIX hp: <http://purl.obolibrary.org/obo/hp#>
SELECT distinct ?report
WHERE {
  ?report a sct:Report .
  ?report sct:referencesConcept* ?radiologicInfiltrate .
  ?radiologicInfiltrate rdfs:subClassOf* hp:0002113 .#radiologic infiltrate
  ?report sct:referencesLocationConcept* ?bronchopulmonarySegment .
  ?bronchopulmonarySegment rdfs:subClassOf* sct:72674008 .#bronshopulmonary segment structure
}
```

Query 3 - Retrieve patient identifier of patients with a radiological finding related to a Widened mediastinum.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
SELECT distinct ?patientId
WHERE {
  ?report a sct:Report .
  ?report sct:referencesToPatient ?patientId .
  ?report sct:referencesConcept* ?widenedMediastinum .
  ?widenedMediastinum rdfs:subClassOf* sct:363646005 .
}
```

Query 4 - Retrieve all reports with radiological findings related to the mediastinum

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
SELECT distinct ?report ?radiologicFinding
WHERE {
  ?report a sct:Report .
  ?report sct:referencesConcept* ?radiologicFinding .
  ?radiologicFinding rdfs:subClassOf* sct:118247008 .
  ?report sct:referencesLocationConcept* ?mediastinum .#mediastinum
  ?mediastinum rdfs:subClassOf* umls:C0025066 .#sct:72410000#
}
```

Query 5 - Retrieve all radiolofy reports related to images with catheter

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
SELECT distinct ?report
WHERE {
  ?report a sct:Report .
  ?report sct:referencesConcept* ?catheter .
  ?catheter rdfs:subClassOf* sct:19923001 .
}
```

Query 6 - Retrieve reports related to fractures in the region of the shoulder

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
PREFIX fma: <http://purl.org/sig/ont/fma/>
SELECT distinct ?report ?fracture
WHERE {
  ?report a sct:Report .
```

```
  ?report sct:referencesConcept* ?fracture .
  ?fracture rdfs:subClassOf* umls:C0016658 .
  ?report sct:referencesLocationConcept* ?shoulder .
  ?shoulder rdfs:subClassOf* fma:25202 .
}
```

Query 7 - Retrieve patient identifiers where nodules of lung were found

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
PREFIX fma: <http://purl.org/sig/ont/fma/>
PREFIX icd10: <http://https://uts-ws.nlm.nih.gov/rest/content/2022AA/source/ICD10AM#>
PREFIX hp: <http://purl.obolibrary.org/obo/hp#>

SELECT distinct ?patientId ?nodulesFinding
WHERE {
  ?report a sct:Report .
  ?report sct:referencesToPatient ?patientId .
  ?report sct:referencesConcept* ?nodulesFinding .
  ?nodulesFinding rdfs:subClassOf* hp:0033608 .
}
```

Query 8 - Retrieve all reports making references to atelectasis and its subtypes

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sct: <http://www.ehealthresearch.no/2022/snomedct-lite#>
PREFIX umls: <https://uts.nlm.nih.gov/uts/umls/concept/>
PREFIX fma: <http://purl.org/sig/ont/fma/>
PREFIX icd10: <http://https://uts-ws.nlm.nih.gov/rest/content/2022AA/source/ICD10AM#>

SELECT distinct ?report ?someTypeOfAtelactasis
WHERE {
  ?report a sct:Report .
  ?report sct:referencesConcept* ?someTypeOfAtelactasis .
  ?someTypeOfAtelactasis rdfs:subClassOf* icd10:J98_1 .
}
```

References

[1] S. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, C. Lehmann, Clinical data reuse or secondary use: Current status and potential future progress, Yearb. Med. Inform. 26 (1) (2017) 38–52, http://dx.doi.org/10.15265/IY-2017-00.

[2] C.B. Forrest, K.M. McTigue, A.F. Hernandez, L.W. Cohen, H. Cruz, K. Haynes, R. Kaushal, A.N. Kho, K.A. Marsolo, V.P. Nair, R. Platt, J.E. Puro, R.L. Rothman, E.A. Shenkman, L.R. Waitman, N.A. Williams, T.W. Carton, PCORnet® 2020: current state, accomplishments, and future directions, J. Clin. Epidemiol. 129 (2021) 60–67, http://dx.doi.org/10.1016/j.jclinepi.2020.09.036.

[3] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers, Stud. Health Technol. Inform. 216 (2015) 574–578, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815923/.

[4] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, D. Madigan, Characterizing treatment pathways at scale using the OHDSI network, Proc. Natl. Acad. Sci. 113 (27) (2016) 7329–7336, http://dx.doi.org/10.1073/pnas.1510502113.

[5] S. Lovestone, EMIF Consortium, The European medical information framework: A novel ecosystem for sharing healthcare data across Europe, Learn. Heal. Syst. 4 (2) (2020) e10214, http://dx.doi.org/10.1002/lrh2.10214.

[6] K.B. Bayley, T. Belnap, L. Savitz, A.L. Masica, N. Shah, N.S. Fleming, Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied, Med. Care 51 (2013) S80, http://dx.doi.org/10.1097/MLR.0b013e31829b1d48.

[7] L. Zhou, L.M. Mahoney, A. Shakurova, F. Goss, F.Y. Chang, D.W. Bates, R.A. Rocha, How many medication orders are entered through free-text in EHRs?–a study on hypoglycemic agents, Annu. Symp. Proc. AMIA Symp. (2012) 1079–1088.

[8] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb. Med. Inform. (2008) 128–144.

[9] P.N. Robinson, Deep phenotyping for precision medicine, Hum. Mutat. 33 (5) (2012) 777–780, http://dx.doi.org/10.1002/humu.22080.

[10] C. Gaudet-Blavignac, V. Foufi, M. Bjelogrlic, C. Lovis, Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: Systematic scoping review, J. Med. Internet Res. 23 (1) (2021) e24594, http://dx.doi.org/10.2196/24594.

[11] I. Coutinho, B. Martins, Transformer-based models for ICD-10 coding of death certificates with portuguese text, J. Biomed. Inform. 136 (2022) 104232, http://dx.doi.org/10.1016/j.jbi.2022.104232.

[12] M.G. Kersloot, F. Lau, A. Abu-Hanna, D.L. Arts, R. Cornet, Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES, J. Biomed. Semant. 10 (1) (2019) 14, http://dx.doi.org/10.1186/s13326-019-0207-3.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS '13, Curran Associates Inc, Red Hook, NY, USA, 2013, pp. 3111–3119.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[15] B. Min, H. Ross, E. Sulem, A.P.B. Veyseh, T.H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2) (2023) 1–40, http://dx.doi.org/10.1145/3605943.

[16] H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun, Pre-trained language models and their applications, Engineering 25 (2023) 51–65, http://dx.doi.org/10.1016/j.eng.2022.04.024.

[17] X. Ho, A.K.D. Nguyen, A.T. Dao, J. Jiang, Y. Chida, K. Sugimoto, H.Q. To, F. Boudin, A. Aizawa, A survey of pre-trained language models for processing scientific text, 2024, arXiv:2401.17824.

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

[19] S.M. Jiménez-Zafra, F. Rangel, M.M.-y. Gómez, Overview of IberLEF 2023: Natural language processing challenges for Spanish and other Iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), Co-Located with the 39th Conference of the Spanish Society for Natural Language Processing, SEPLN 2023, CEURWS. Org, 2023.

[20] L. Chiruzzo, S.M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: natural language processing challenges for Spanish and other Iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), Co-Located with the 40th Conference of the Spanish Society for Natural Language Processing, SEPLN 2024, CEUR-WS. Org, 2024.

[21] M. Rojas, J. Dunstan, F. Villena, Clinical flair: A pre-trained language model for Spanish clinical natural language processing, in: Proceedings of the 4th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, WA, 2022, pp. 87–92, http://dx.doi.org/10.18653/v1/2022.clinicalnlp-1.9, URL https://aclanthology.org/2022.clinicalnlp-1.9.

[22] M.A. Shaaban, A. Akkasi, A. Khan, M. Komeili, M. Yaqub, Fine-tuned large language models for symptom recognition from Spanish clinical text, 2024, arXiv:2401.15780.

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

[24] C.P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199, http://dx.doi.org/10.18653/v1/2022.bionlp-1.19.

[25] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355, http://dx.doi.org/10.18653/v1/W18-5446.

[26] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 58–65, http://dx.doi.org/10.18653/v1/W19-5006.

[27] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797, http://dx.doi.org/10.1016/j.media.2020.101797.

[28] M.d.l.I. Vayá, J.M. Saborit, J.A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, J.M. Salinas, BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, 2020, http://dx.doi.org/10.48550/arxiv.2006.01174, arXiv:2006.01174.

[29] J.C.Y. Seah, C.H.M. Tang, Q.D. Buchlak, X.G. Holt, J.B. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J.F. Lambert, B. Hachey, S.J.F. Hogg, B.P. Johnston, C. Bennett, L. Oakden-Rayner, P. Brotchie, C.M. Jones, Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study, Lancet Digit. Heal. 3 (8) (2021) e496–e506, http://dx.doi.org/10.1016/S2589-7500(21)00106-0.

[30] A.J. DeGrave, J.D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, Nat. Mach. Intell. 3 (7) (2021) 610–619, http://dx.doi.org/10.1038/s42256-021-00338-7.

[31] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, vol. 3180, Bologna, Italy, 2022, pp. 179–203.

[32] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (2) (1994) 161–174.

[33] A.R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, Proc. AMIA Symp. (2001) 17–21.

[34] L. Hammami, A. Paglialonga, G. Pruneri, M. Torresani, M. Sant, C. Bono, E.G. Caiani, P. Baili, Automated classification of cancer morphology from Italian pathology reports using natural language processing techniques: A rule-based approach, J. Biomed. Inform. 116 (2021) 103712, http://dx.doi.org/10.1016/j.jbi.2021.103712.

[35] Projecte de medicina personalitzada big data | educational, 2023, URL https://www.san.gva.es/es/web/investigacio/med-p-big-data-. (Accessed: 04 Jan 2023).

[36] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Heal. 3 (1) (2021) 1–13, http://dx.doi.org/10.1145/3458754.

[37] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[38] L. Marco-Ruiz, C. Pedrinaci, J.A. Maldonado, L. Panziera, R. Chen, J.G. Bellika, Publication, discovery and interoperability of clinical decision support systems: a linked data approach, J. Biomed. Inform. 62 (2016) 243–264.

[39] R. Arp, B. Smith, A.D. Spear, Building Ontologies with Basic Formal Ontology, The MIT Press, 2015.

[40] Schema.org - schema.org, 2023, URL https://schema.org/. (Accessed: 04 Jan 2023).

[41] Dublin core metadata initiative (DCMI), 2023, URL https://www.dublincore.org/. (Accessed: 04 Jan 2023).

[42] C. Pedrinaci, J. Domingue, Toward the next wave of services: Linked services for the web of data, J. UCS 16 (13) (2010) 1694–1719.
[43] FAIRsharing | educational, 2023, URL https://fairsharing.org/educational. (Accessed: 04 Jan 2023).
[44] European open science cloud (EOSC) | shaping Europe's digital future, 2023, URL https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud. (Accessed: 04 Jan 2023).
[45] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, J. Biomed. Inform. 87 (2018) 12–20, http://dx.doi.org/10.1016/j.jbi.2018.09.008.
[46] S. Datta, Y. Si, L. Rodriguez, S.E. Shooshan, D. Demner-Fushman, K. Roberts, Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning, J. Biomed. Inform. 108 (2020) 103473, http://dx.doi.org/10.1016/j.jbi.2020.103473.
[47] C. Tao, M. Filannino, Ö. Uzuner, Prescription extraction using CRFs and word embeddings, J. Biomed. Inform. 72 (2017) 60–66, http://dx.doi.org/10.1016/j.jbi.2017.07.002.
[48] M. Oleynik, D.F.C. Patrão, M. Finger, Automated classification of semi-structured pathology reports into ICD-O using SVM in portuguese, Stud. Health Technol. Inform. 235 (2017) 256–260.
[49] J.R. Mitchell, P. Szepietowski, R. Howard, P. Reisman, J.D. Jones, P. Lewis, B.L. Fridley, D.E. Rollison, A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT network): Development study, J. Med. Internet Res. 24 (3) (2022) e27210, http://dx.doi.org/10.2196/27210.
[50] A. Rios, E.B. Durbin, I. Hands, R. Kavuluru, Assigning ICD-O-3 codes to pathology reports using neural multi-task training with hierarchical regularization, ACM Conf. Bioinform., Comput. Biol., Health Inform. (ACM BCB) 32 (2021) 32, http://dx.doi.org/10.1145/3459930.3469541.
[51] P. López-Úbeda, M.C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L.A. Ureña-López, M.T. Martín-Valdivia, Automatic medical protocol classification using machine learning approaches, Comput. Methods Programs Biomed. 200 (2021) 105939, http://dx.doi.org/10.1016/j.cmpb.2021.105939.
[52] A. Olthof, P. Shouche, E. Fennema, F. IJpma, R. Koolstra, V. Stirler, P. van Ooijen, L. Cornelissen, Machine learning based natural language processing of radiology reports in orthopaedic trauma, Comput. Methods Programs Biomed. 208 (2021) 106304, http://dx.doi.org/10.1016/j.cmpb.2021.106304.
[53] D. Karlsson, M. Nyström, R. Cornet, Does SNOMED CT post-coordination scale? Stud. Health Technol. Inform. 205 (2014) 1048–1052.
[54] J.M. Banda, Fully connecting the observational health data science and informatics (OHDSI) initiative with the world of linked open data, Genom. Inform. 17 (2) (2019) e13, http://dx.doi.org/10.5808/GI.2019.17.2.e13.
[55] J.M. Banda, Y. Halpern, D. Sontag, N.H. Shah, Electronic phenotyping with APHRODITE and the observational health sciences and informatics (OHDSI) data network, AMIA Jt. Summits Transl. Sci. Proc. AMIA J. oint Summits Transl. Sci. 2017 (2017) 48–57.
[56] V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, T.I. Leung, E.P. Minty, T.E. Sweeney, E. Gyang, N.H. Shah, Learning statistical models of phenotypes using noisy labeled training data, J. Am. Med. Inform. Assoc.: J. AMIA 23 (6) (2016) 1166–1173, http://dx.doi.org/10.1093/jamia/ocw028.
[57] N. Viani, R. Botelle, J. Kerwin, L. Yin, R. Patel, R. Stewart, S. Velupillai, A natural language processing approach for identifying temporal disease onset information from mental healthcare text, Sci. Rep. 11 (1) (2021) 757, http://dx.doi.org/10.1038/s41598-020-80457-0.
[58] E.T. Sholle, M. Cusick, M.A. Davila, J. Kabariti, S. Flores, T.R. Campion, Characterizing basic and complex usage of i2b2 at an academic medical center, AMIA Jt. Summits Transl. Sci. Proc. (2020) 589–596.
[59] RadLex term browser, 2024, URL https://radlex.org/. (Accesssed: 24 Jan 2024).
[60] J.-M. Papaioannou, P. Grundmann, B. van Aken, A. Samaras, I. Kyparissidis, G. Giannakoulas, F. Gers, A. Loeser, Cross-lingual knowledge transfer for clinical phenotyping, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 900–909, URL https://aclanthology.org/2022.lrec-1.95.
[61] A. Alekseev, Z. Miftahutdinov, E. Tutubalina, A. Shelmanov, V. Ivanov, V. Kokh, A. Nesterov, M. Avetisian, A. Chertok, S. Nikolenko, Medical crossing: a cross-lingual evaluation of clinical entity linking, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4212–4220, URL https://aclanthology.org/2022.lrec-1.447.
[62] Y. Lin, P. Hoffmann, E. Rahm, Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models, SN Comput. Sci. 3 (2022) 387, http://dx.doi.org/10.1007/s42979-022-01295-7.
[63] T. Lai, H. Ji, C. Zhai, BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1631–1639, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.140.
[64] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4228–4238, http://dx.doi.org/10.18653/v1/2021.naacl-main.334.
[65] D. Agarwal, R. Angell, N. Monath, A. McCallum, Entity linking via explicit mention-mention coreference modeling, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4644–4658, http://dx.doi.org/10.18653/v1/2022.naacl-main.343.