PADCHEST: A LARGE CHEST X-RAY IMAGE DATASET WITH MULTI-LABEL ANNOTATED REPORTS

A PREPRINT

Aurelia Bustos
Department of Software and Computing Systems
University Institute for Computing Research
University of Alicante, Spain
aurelia@medbravo.org

Antonio Pertusa
Department of Software and Computing Systems
University Institute for Computing Research
University of Alicante, Spain
pertusa@ua.es

Jose-Maria Salinas
Department of Health Informatics
Hospital San Juan de Alicante, Spain
salinas_josser@gva.es

Maria de la Iglesia-Vayá
Centre of Excellence in Biomedical Image
Regional Ministry of Health, Valencia, Spain
delaiglesia_mar@gva.es

February 8, 2019

ABSTRACT

We present a labeled large-scale, high resolution chest x-ray dataset for the automated exploration of medical images along with their associated reports. This dataset includes more than 160,000 images obtained from 67,000 patients that were interpreted and reported by radiologists at Hospital San Juan Hospital (Spain) from 2009 to 2017, covering six different position views and additional information on image acquisition and patient demography. The reports were labeled with 174 different radiographic findings, 19 differential diagnoses and 104 anatomic locations organized as a hierarchical taxonomy and mapped onto standard Unified Medical Language System (UMLS) terminology. Of these reports, 27% were manually annotated by trained physicians and the remaining set was labeled using a supervised method based on a recurrent neural network with attention mechanisms. The labels generated were then validated in an independent test set achieving a 0.93 Micro-F1 score. To the best of our knowledge, this is one of the largest public chest x-ray database suitable for training supervised models concerning radiographs, and the first to contain radiographic reports in Spanish. The PadChest dataset can be downloaded from http://bimcv.cipf.es/bimcv-projects/padchest/.

Keywords X-ray image dataset · Deep neural networks · Radiographic findings · Differential diagnoses · Anatomical Locations

1 Introduction

Chest x-rays are essential for both the screening and the diagnosis of pulmonary, cardiovascular, bone and other thoracic disorders. The adequate interpretation of the radiographic findings requires medical training acquired over many years, with radiologists being the most qualified professionals in this fields. Due to increasing workload pressures, many radiologists today have to read more than 100 x-ray studies daily. Therefore, automated tools trained to predict the risk of specific abnormalities given a particular x-ray image have the potential to support the reading workflow of the radiologist. Those tools could be used to enhance the confidence of the radiologist or prioritize the reading list where critical cases would be read first. Decision support systems (DSS) designed as tools to assist in the clinical interpretation of chest x-rays would therefore fulfill an unmet need.

Deep learning techniques are currently obtaining promising results and perform extremely well in a variety of sophisticated tasks [1], especially those related to computer vision, often equaling or exceeding human performance [2]. The application of deep neural networks to medical imaging and chest radiographs in particular, has become a growing

area of research in recent years [3]. For instance, [4] trained a Convolutional Neural Network (CNN) to classify and localize 8 pathologies using the chest x-ray database (*ChestX-Ray8*) which comprised 108,948 frontal-view x-ray images of 32,717 different patients. Using the same repository, [5] extended the annotations to 14 different pathologies (*ChestX-Ray14*) and designed a model with a deeper CNN architecture to classify images as 14 pathological entities. This method was reported to obtain greater diagnostic efficiency in the detection of pneumonias when compared to that of radiologists. [6] proposed the attention guided CNN to help combine global and local information in order to improve recognition performance. Chest-XRay14 was also employed by [7] to design a network architecture that combines text and image with attention mechanisms capable of generating text that describes the image, while [8] introduced a hierarchical model of Recurrent Neural Networks (RNN) with which to generate long paragraphs from the images and obtain semantically linked phrases for the same purpose.

Despite claims that they achieve and/or surpass physician-level performance, current deep learning models for the classification of pathologies using chest x-rays are proving not to be generalizable across institutions and not yet ready for adoption in real-world clinical settings [9]. Moreover, warnings of potential unintended consequences of their use are discussed by [10].

It is unclear how to extend the significant success in computer vision using deep neural networks to the medical domain [11]. Open questions concerning medical radiology datasets that still need to be addressed are:

- How to annotate the huge amount of medical images required by deep learning models [11] and meet the required quality. Large-scale crowd-sourced hand-annotation which has proved successful in the general domain, e.g. *ImageNet* [12], is not feasible because of the medical expertise required to carry it. This, is compounded by the fact that the semantic interpretation and extraction of medical knowledge from the corpora of medical text in unstructured natural language remains a challenge [13] and is an area of active research [14].

- The clinically relevant image labels that need to be defined and which criteria should be followed to annotate them [11].

- How to deal with uncertainties in radiology texts. Medical data is characterized by uncertainty and incompleteness and machine learning decision support systems (ML-DSS) need to adapt to input data reflecting the nature of medical information, rather than at imposing an idea of data accuracy and completeness that does not fit patient records and medical registries, for which data quality is far from optimal. In this respect, [10] advises caution as regards the unintended consequences of adopting ML-DSS that demise context and ignore the fact that observer variability obeys not only interpretative deficiencies but also intrinsic variability in the observed phenomena.

- How to effectively control for potential confounding factors, such as the presence of tubes, catheters, the quality of the image as assessed by radiologists, patient position, etc. and unbalanced entity prevalence, which models learn to exploit as predictive features to the detriment of clinical radiological patterns.

There are a number of publicly available chest x-ray datasets that can be used for image classification and retrieval tasks. The National Institute of Health of America (NIH) repository [4] contains 112,120 frontal-view chest x-rays, corresponding to 30,805 different patients, and multi-labeled with 14 different thoracic diseases [5]. The Korean Institute of Tuberculosis dataset [15] consists of 10,848 DICOMs, of which 3,828 show tuberculosis abnormalities. The Indiana University dataset [16] comprises 7,470 frontal and lateral chest x-ray images, corresponding to 3,955 radiology reports with disease annotations, such as cardiac hypertrophy, pulmonary edema, opacity, or pleural effusion. The JSRT dataset [17] consists of 247 x-rays of which 154 have been labeled for lung nodules (100 malignant). [18] also provides masks of the lung area for the evaluation of segmentation performance. The Shenzhen dataset [19] has a total of 662 images belonging to two categories (normal and tuberculosis).

With regard to the annotation methods applied, [20, 11] used a large dataset that included 780,000 documents from 216,000 images comprising CT, MR, PET and other image modalities, and automatically mined categorical semantic labels using a non-parametric topic modeling method. The resulting annotations were judged to be non-specific. In order to increase disease specificity, the authors have matched frequent pathology types using a disease ontology and semantics. This method, however, assigned specific disease labels to around only 10% of the dataset. In *ChestX-Ray-8*, [4] used MetaMap [21], DNorm [22] and custom negation rules applied to a syntactic parser in order to label the presence or otherwise of 8 entities, which was further expanded to 14 entities in 125,000 images (*Chest-XRay14*), also by applying analogous methods. They validated the image labeling procedure against 3,800 annotated reports of x-ray images from OpenI - Indiana DB- [16].

Outstanding questions, such as those mentioned above, still remain unaddressed by the medical image datasets currently available. This problem is compounded by the fact that the, medical annotations used as a ground-truth are reduced to a small number of entities and, on many occasions, owing to the inherent limitations of the natural language processing

(NLP) techniques applied to automate their extraction, these annotations contain omissions, inconsistencies and are not validated by physicians.

In this work, we propose a dataset called PadChest (PAthology Detection in Chest radiographs) which is, to the best of our knowledge, one of the largest and most exhaustively labeled public chest x-ray dataset and the only one to contain the source report excerpts, which are written in Spanish. The labels of the images are mapped onto standard Unified Medical Language System (UMLS) terminology and can therefore be used regardless of the language. They cover the full spectrum of thoracic entities, contrasting with the much more reduced number of entities annotated in previous datasets. Moreover, entities are localized with anatomical labels, images are stored in high resolution, and extensive information is provided, including the patient's demography, type of projection and acquisition parameters for the imaging study, among others.

In contrast to previous approaches that relied solely on automatic annotation tools, in this work the labeling required for the baseline ground-truth was carried out manually by trained physicians. While tools to annotate medical text in Spanish are less extensively developed and tested than in English, the most compelling reason for opting for an annotation carried out manually by physicians was to maximize the reliability of the baseline ground-truth.

Using a laboriously developed ground-truth as a baseline, we propose a supervised annotation method based on deep neural networks and labeling resolution rules in order to extract labels from the remaining reports (73% of the samples). This pipeline is designed for large-scale exhaustive annotation of Spanish chest x-ray reports and aims to overcome some common NLP challenges in the medical domain. For instance, the proposed method deals with anaphora resolution in which findings are mentioned in different sentences or even in different studies for the same patient and whose interpretation depends upon another expression in that context. This method also deals with co-reference resolution, in which words for locations are related to the entities that they are describing, and also with hedging statements, in which uncertainty, probability and indirect expressions are to be learned by deep neural models.

Moreover, in an effort to incorporate uncertainty into the ground-truth, we differentiate radiographic findings from differential diagnoses, acknowledging the existence of two distinct image classification problems. Radiographic findings are completely observable in the images and we, therefore hypothesize that state-of-the-art deep learning models could achieve reliable results for this task. However, differential diagnoses are characterized by intrinsic uncertainty and a highly multidimensional context which is not included in the image. The expectation that automated methods could obtain similar results for differential diagnoses to those of doctors using only x-rays as input may, therefore, be unrealistic. This is because diagnoses such as "pneumonia" are not based solely on the x-ray image and depend to a great extent on external factors, such as laboratory tests, physical examinations, symptoms and signs, temporal clues and clinical judgment, which may additionally vary among practitioners.

The main objectives of the proposed dataset are:

- To broaden the scope of radiographic diagnoses and findings that trainable models can learn from chest x-ray images, including respiratory, cardiac, infectious, neoplastic, bone and soft tissue diagnoses, the positions of nasogastric tubes, endotracheal tubes, central venous catheters, electrical devices, etc.

- To increase the disease detection capabilities of trained models by providing the localizations of the entities as regions of interest.

- To make available all relevant metadata, along with all the entities described by radiologists, so as to help control potential confounders in predictive models. For example projection types that dictate the adequate interpretation of radiographic findings and patient positioning have been identified as confounding factors in deep learning models. Indeed, the first task for radiologists is to identify the type of projection before reading the x-ray in order to correctly interpret the findings.

- To help advance automatic clinical text annotations in Spanish. To the best of our knowledge, this is the first publicly available chest x-ray dataset containing excerpts from radiology reports in this language. Although the excerpts from the report are provided in Spanish, the labels are mapped onto biomedical vocabulary unique identifier (CUIs) codes, thus making the dataset usable regardless of the language.

- Training text models with the proposed deep learning methods would make it possible to help automatically label other large-scale x-ray repositories in Spanish speaking health institutions, using the source code provided.

The remainder of the paper is as follows. Section 2 describes the methodology employed to build the dataset, including the manually labeled subset and the automatic labeling of the remaining reports using deep neural networks. Section 3 shows the evaluation results of the automatic-labeling methods described in the previous section, while Section 4 details the statistics of the dataset. Finally, Section 5 addresses the discussion, conclusions and future work.
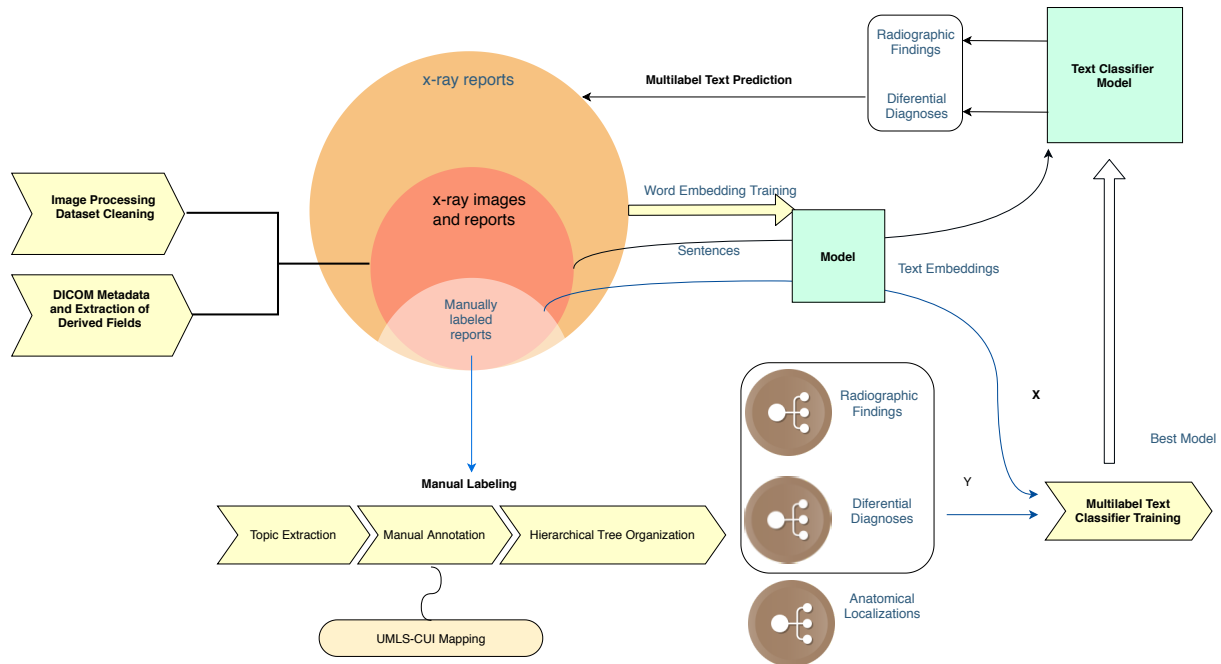
Figure 1: Dataset Building Pipeline: The PadChest dataset consists of 206,222 x-ray reports (large circle), 109,931 (middle circle) of which had their 160,868 corresponding images in DICOM format and were acquired from the years 2009 to 2017. A subsample of 27,593 reports (pale oval region) from the years 2014 to 2017 were manually labeled and further used to train a multi-label text classifier based on neural networks in order to annotate the remaining dataset with radiographic findings and differential diagnoses. Anatomical localizations were extracted using regular expressions. Different diagnoses, radiographic findings and anatomical localizations were mapped onto the NLM Unified Medical Language System (UMLS) controlled biomedical vocabulary unique identifiers (CUIs) and organized into semantic concept trees.

2 Material and methods

The PadChest dataset consists of all the available chest-x rays that had been interpreted and reported by 18 radiologists at the *Hospital Universitario de San Juan, Alicante* (Spain) from Jan 2009 to Dec 2017, amounting to 109,931 studies and 168,861 different images, as shown in Tab. 7.

This project was approved by the institutional research committee, and both the images and the associated reports were made anonymous and de-identified by the Medical Image Bank of the Valencian Community at the Department of Universal Health and Public Health Services (BIMCV-CSUSP) and the Health Informatics Department at San Juan Hospital.

The PadChest dataset can be downloaded from the repository of the medical imaging bank (BIMCV - PADCHEST[1]), enabled by the Medical Image Bank of the Valencian Community (BIMCV). The BIMCV has launched various projects regarding population medical images, whose objective is to develop and implement an infrastructure with a massive storage capacity following the R&D Cloud CEIB architecture [23]. One of the missions of this bank is to promote the publication of scientific knowledge as open data by its affiliated health institutions.

PadChest contains image files adding up to 1 TB, a csv file with 33 fields for each study and an instruction file containing field descriptions, examples and search information for efficient image retrieval. An example of a dataset study with two projections can be found in C, along with its associated labels and additional information fields.

The methodology employed to build PadChest comprises the following main steps:

- Pre-processing of the images and DICOM metadata extraction.

- Pre-processing of the reports.

---

[1] http://bimcv.cipf.es/bimcv-projects/padchest/

4

(a) P-A   (b) Lateral   (c) Lordotic   (d) A-P supine   (e) A-P

(f) P-A   (g) Lateral   (h) Lordotic   (i) A-P supine   (j) A-P
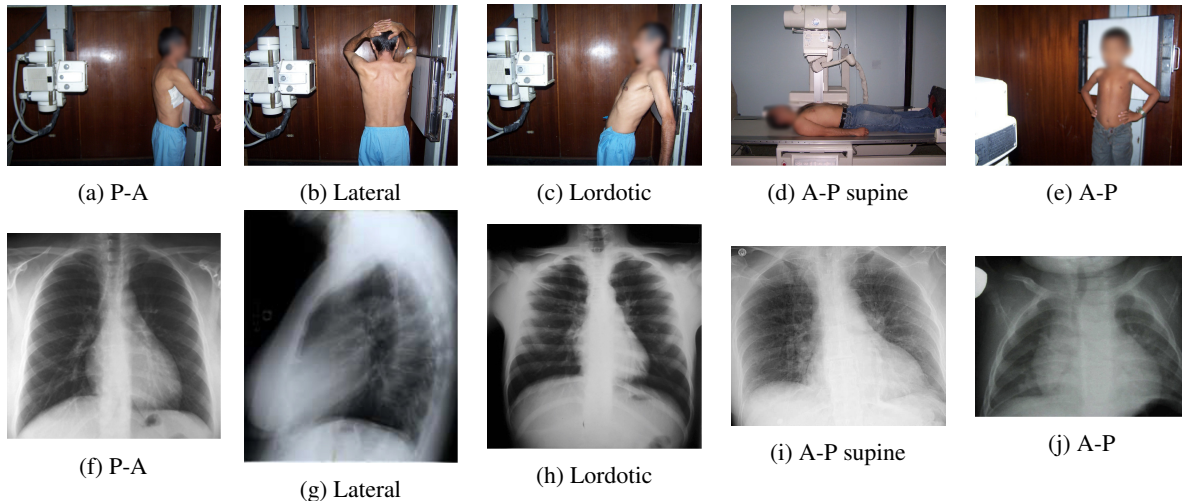
Figure 2: Common chest x-ray projections.

- Manual medical annotations using a hierarchical taxonomy of radiographic findings, differential diagnoses and their anatomic locations.

- Automatic labeling of the remaining studies.

2.1 Pre-processing of the images and DICOM metadata extraction

The images were processed by rescaling the dynamic range using the DICOM window width and center, when available. They were not resized to avoid the loss of resolution.

Some images were initially excluded when: 1) DICOM pixel data were not readable; 2) the photometric interpretation was missing or MONOCHROME1; 3) the modality was missing or it was Computed Tomography (CT) rather than x-rays; 4) the projections were lateral horizontal, oblique or trans-thoracic; 5) the anatomic image protocol was other than that of the chest (e.g humerus, abdomen, ..), and 6) the study report was missing or the radiography interpretation was not identifiable.

In addition to the standard DICOM metadata, the information on image projection and radiographic positioning was extracted. These data were found in different DICOM fields in non-structured free text (Position View, Patient Orientation, Series Description and Code Meaning). The projections principally identified after excluding obliques were antero-posterior (AP), postero-anterior (PA) and lateral (L). The different body positions were erect, either standing or sitting, decubitus or lying down, supine or lying on back, lateral decubitus right lateral and left lateral. In addition, different protocols were identified based on different clinical scenarios: standard standing PA and L, mobile x-ray (for patients unable to stand) in either AP erect in bed or AP supine, pediatric protocol for patients up to 3 years old, lordotic views, ribs and sternum modality views.

The projection information is highly relevant for diagnosis. For example, AP views, which are commonly used in pediatric patients, show an enlarged heart silhouette (Fig. 2j) that should not be interpreted as cardiomegaly, but merely the expected large-depth ratio of reversed organ observation (Fig. 3). Another illustrative example is the distinct pattern that pleural effusions have in the standing position (Fig. 4a), in which a typical meniscus sign is commonly found as opposed to decubit projections (Fig. 4b). Given that the number of different projections is unbalanced (for instance, PA followed by lateral projections typically comprise the majority of chest x-rays), there is the risk that none of the other projections will have sufficient instances with which to train models capable of discriminating pathological from non-pathological patterns in the context of the projection.

There are particular radiological landmarks that differentiate projections, which radiologists are trained to identify. For instance, in the case of PA projections, these landmarks are the presence of air in the gastric chamber and the scapulae projected outside the lung fields. Although these features can be learned, models trained in unbalanced datasets with a poor representation for different projections may not have sufficient instances to properly learn those patterns. An illustrative example is when the heart enlargement in AP projections is attributable only to the effect of the projection, while the trained model erroneously predicts cardiomegaly.
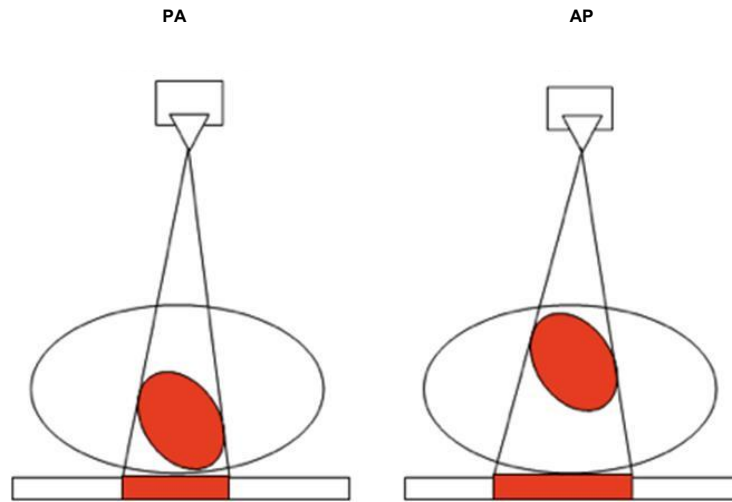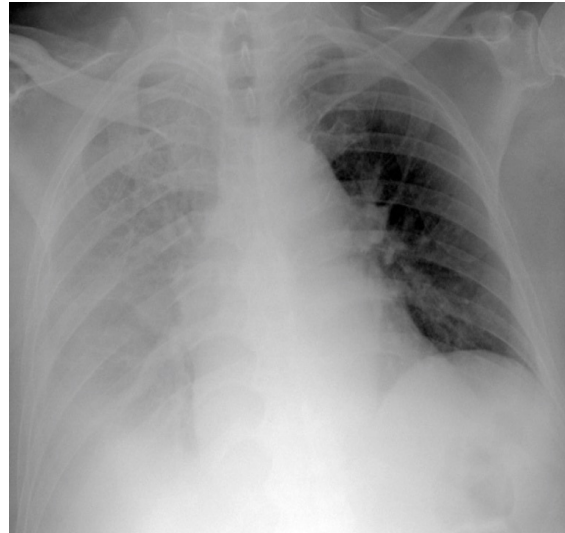
Figure 3: Heart size projection in Postero-Anterior (PA) vs Antero-Posterior (AP). The projection of the heart (red silhouette) illustrates that anatomical ratios depend on the plane distance to the x-ray source.

(a) P-A

(b) Decubit

Figure 4: Pleural effusion in different projections: A bipedestation projection (a) shows the meniscus sign in which the fluid accumulates in the subpulmonary region, ascends through the thoracic wall and through the paramedian zone. In decubit projection (b) there is no meniscus sign. As the liquid goes to the most declining area there is a diffuse increase in hemithorax density and a loss of the net limit of the diaphragm with occupation of the pulmonary vertex by apical cap, costo-phrenic angle blunting and a thickening of the smaller fissure.

Given that there were numerous combinations of projections, positions and protocols and all this information was not uniformly reported in DICOM meta-data, we decided to group them into 6 main classes: standard PA, standard L, AP vertical or erect, AP horizontal or supine, pediatric, and rib views.

For those images without DICOM information on the type of projection (20,367 samples), we used a pre-trained model available at `https://github.com/xinario/chestViewSplit` and implemented a custom method to preprocess and load PadChest images in the expected format. This method uses a pre-trained ResNet-50 [24] CNN model initialized with the ImageNet [12] weights and trained with fine-tuning on chest x-rays. Note that these automatically labeled projection samples should not subsequently be used to train another classifier with radiographs as inputs and projections as outputs. These subsets of projections were the only ones obtained from the images using an automatic classifier, unlike all the other data provided in PadChest, which can be used to train any classifier on input images, as they were obtained from the reports.

2.2 Preprocessing of the reports

The text dataset consisted of 206,222 study reports. There was only one report for each study, and if a study had two projections, such as PA and L, then the radiography description, therefore, included both results in the same report. Not all study reports had their x-ray images available, and conversely not all available x-rays had corresponding reports.

The ground truth was obtained from raw text reports but was not usable "as is" for the following reasons:

- The reports were in unstructured free-text and did not use standardized dictionaries or medical codes.

- Although the radiography description was consistently present in the report, other sections, such as patient's history, study reason, and diagnostic impression, were not systematically included in the report and, when present, those sections where not clearly delimited or did not follow a consistent pattern. Particularly, content format largely varied over the years.

- In several cases, the radiography description contained other projection types, such as sinus and abdominal types, whose interpretation was included together with the chest x-rays in the same report.

The following pipeline was consequently implemented[2] in order to obtain a final curated labeled dataset:

1. Lowercased text and removal of accents.

2. The identification and extraction of the text containing the radiography description of chest x-rays in each report was carried out using an iterative incremental approach based on regular expressions and a repeated random sampling review to improve and/or add more regular expressions. A total of 167 (0.6%) reports were excluded because the radiography description was not retrievable with the chosen expressions or because the section was empty.

3. Removal of non alpha-numerical characters excluding dots and spaces.

4. Removal of stopwords, with the exception of ['sin', 'no', 'ni', 'con'], using Spanish stopword removal from NLTK [25].

5. Spanish stemming to remove morphological affixes from words, leaving only word stems using NLTK.

The complete report dataset had 501,840 sentences, amounting to 3.8 million words. The vocabulary size before preprocessing was 14,234 unique words, and was downsized to 9,691 different tokens after stopword removal (310 words) and stemming.

The mean number of tokens per sentence was 7.1 and the median was 5 tokens per sentence.

2.3 Medical codes annotation

A total of 27,593 reports were manually annotated following a hierarchical taxonomy described below. The manually labeled dataset was subsequently used to train and validate a multi-label text classifier, thus enabling the automatic annotation of the remaining 82,338 reports.
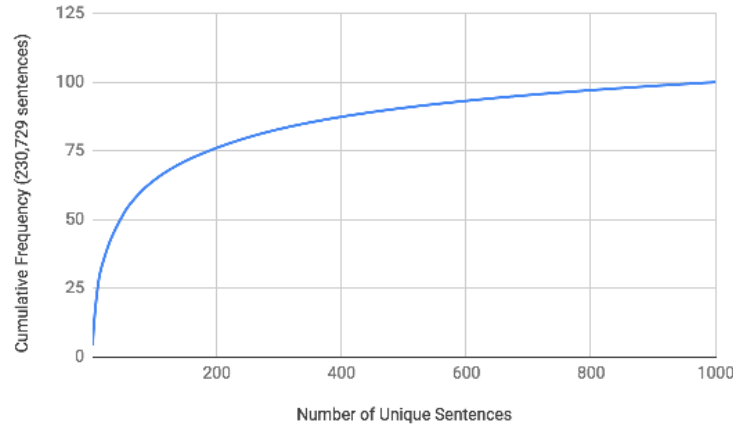
Figure 5: Pareto curve of sentence redundancy. The cumulative frequency of the 1,000 most repeated sentences in the entire set of x-ray reports amounted to 230,729 sentences (the total number of sentences in the entire dataset was 500,000).

2.3.1 Topic extraction

In order to help the practitioner label the dataset with radiographic findings and diagnoses, an automatic process was performed on the pre-processed reports so as to extract a series of topics (concepts) that were subsequently used to speed-up the manual labeling process.

For this goal, first, the preprocessed radiography reports were split into sentences and unique sentences were ordered by frequency.

Of 145,134 unique sentences from a total of 500,000 sentences, the 1,000 first most repeated ones comprised 230,729 (46%). Fig. 5 shows a Pareto curve of the 1,000 most frequent sentences. Note that 20% of them covered up to 76% of 230,729 sentences. In order to exploit this redundancy, the strategy employed was to manually label at sentence level.

Topic extraction discovers keyphrases and concepts in each sentence based on the frequency and linguistic patterns in the text. In order to facilitate the manual annotation task, the topics of sentences were automatically extracted and then presented to physicians for evaluation. For example, the list of phrases from the topic associated with the Chronic obstructive pulmonary disease (COPD) is: "epoc sign radiolog atrap aere sugest marc dorsal import inter cardiomegalia aereo lev significativos prominent vascular apical torax escoliosis derecho".

Three different unsupervised clustering methods with which to extract the topics from the reports were evaluated:

1. Non-negative Matrix Factorization (NMF), [26]. First, term frequency–inverse document frequency (tf-idf) statistics were extracted to reflect how important a word is in the corpus.An NMF model was then fitted with the tf-idf features using two different objective functions: the Frobenius norm, and the generalized Kullback-Leibler divergence.

2. Latent Dirichlet Allocation (LDA), [27]. In this case, tf (raw term frequency) features were extracted to fit an LDA model.

3. k-Means. First, Doc2Vec [28] representations were extracted, which is a generalization of Word2Vec [29] for learning vectors from documents. The following parameters were used to learn the representations: Learning rate 0.025, size of word vectors 300, size of the context window 10, number of epochs 55, minimum . number of word occurrences 5, number of negative sampled 5, loss function negative sampling, sampling threshold 10^{-3}. Once the representations had been extracted, the k-Means algorithm was used to cluster these vectors into 20 topics as shown in Tab. 1.

The output from these methods is a list of topics, each represented as a list of terms. The results were evaluated qualitatively for semantic performance, and homogeneity, and the 20 topics obtained with Doc2Vec and k-Means were chosen by physician consensus. Grouping sentences by topic increased efficiency, thus allowing physicians to find batches of sentences, in which the same labels could be propagated seamlessly, more easily.

---

[2]Source code available at `https://github.com/auriml/Rx-thorax-automatic-captioning/blob/master/report_preprocessor.ipynb`

Topic	Semantic group
1	hypoexpansion basal, laminar atelectasis, bronchiectasis, pleural effusion ,pneumonia, infiltrates
2	pleural effusion both sides, bronchovascular markings, gynecomastia, interstitial pattern bilateral, pulmonary edema, respiratory distress, heart insufficiency
3	tracheostomy tube, endotracheal tube, NSG tube, chest drain tube
4,12	exclude sentences without radiolographical findings mainly mentioning additional imaging test recommended in the follow-up
5	scoliosis, kyphosis, vertebral degenerative changes, osteopenia, osteoporosis
6	COPD signs, air trapping, chronic changes, hyperinflated lung both sides, flattened diaphragm, emphysema
7	vertebral compression
8	NSG tube, hiatal hernia
9,15,16, 18	normal study
10	calcified granuloma, nipple shadow
11	callus rib fracture, humeral prosthesis, osteosynthesis material, cervical rib
13	unchanged
14	central venous catheter, reservoir central, pacemaker
17	cardiomegaly, vascular hilar enlargement, pulmonary hypertension, hilar congestion, vascular redistribution
19	surgery, sternotomy, diaphragmatic eventration, hemidiaphragm elevation, costophrenic angle blunting

Table 1: Sentence topics: Sentences were automatically assigned to 20 topics with the sole purpose of organizing unique sentences by semantic logic. This would, therefore serve as a tool with wich to help the physician in the manual extraction of labels. Each of these topics captured different medical semantic groups that facilitated the manual task of label extraction by ordering the semantically close sentences.

2.3.2 Manual labeling

A total of 22,120 unique sentences corresponding to 27,593 study reports were labeled and manually reviewed by trained physicians. Medical entities were extracted as per physician criteria and mapped onto Unified Medical Language System (UMLS [30]) controlled biomedical vocabulary unique identifiers. Medical entities whose exact meaning could not be found in any term of the UML metathesaurus, but that were relevant for annotation purposes were also extracted using the physician's labels but without assigning a code to them. Four additional labels with special coding rules were explicitly defined in order to maintain consistency in the following cases:

- Sentences describing normality, i.e. those that either reported complete resolution, did not describe radiographic findings or negated their presence were labeled as "Normal".

- Sentences that qualified images as suboptimal or with deficient technique as reported by radiologists were labeled as "Suboptimal".

- Sentences that were not interpretable or that did not mention any radiographic findings or diagnoses were labeled as "Exclude".

- Sentences mentioning that there were no changes in a medical entity described previously but not included in the report were labeled as "Unchanged" unknown entity. This procedure allows to distinguish pathologic from normal studies even when the medical entity is not mentioned and only a temporal reference is given. This case is frequent on temporal series of studies done on hospitalized patients to control the evolution of a pathological finding. It is important to remark that before training a model to predict pathologies from images using PadChest, labels of type "Unchanged" are intended to be replaced by labels reported for the same patient, if available in a prior study. Otherwise, the label should be removed as it could not be learned from the image.

All the sentences were assigned to multiple labels or at least one. When a list of potential diagnoses was reported, all differential diagnoses were labeled. Similarly, when different radiographic findings were described all of them were labeled.

As a result, both the sentences and hence the reports (as a sequence of sentences) were multi-label. Finally, the following consecutive rules were applied to ensure consistency when obtaining the final unique set of multi-labels for each report:

1. The set of unique labels was obtained as the union operation of all labels assigned to each report.

2. The report was assigned a single label "Normal" only if in addition to that label, there were no other labels regarded as a radiographic finding or diagnosis.

3. The report was assigned a single label "Exclude" if this was the only label.

2.3.3 Organization of labels in hierarchical taxonomies

The list of unique manually annotated labels was organized into three hierarchical trees (see A) in order to facilitate the exploitation and retrieval of images by semiological groups, differential diagnoses and anatomic locations. The hierarchical organization criteria and content of the three resulting taxonomy trees were reviewed by a radiologist with 20 years of clinical experience.

The main purpose of the hierarchic organization is to allow comprehensive retrieval of the studies grouped by higher hierarchical levels, thus enabling the construction of partitions using different criteria when compiling training sets for machine learning techniques. This is particularly relevant in order to control balance and granularity when deciding the classes to be inferred by classification methods.

The hierarchies are multi-axial, that is, the same term can be found in different branches, and child-parent relationships should satisfy that a child entity "is a" parent entity. For instance, "chronic tuberculosis" is "tuberculosis". Following these criteria, each medical concept was classified and assigned a node of the radiographic finding versus the differential diagnosis tree, and spatial concepts were assigned a node in the anatomical locations tree.

The criteria applied to each of the three hierarchical trees were:

- Radiographic Findings Tree (A.1.1): Findings were defined as any medical entity or diagnosis that radiologists can assert based solely on the interpretation of a chest x-ray image without any additional knowledge of the patient's clinical information. Medical entities that can be directly diagnosed by interpreting the image are, for example, those affecting the bones (e.g, fractures, degenerative diseases, etc) or those involving alterations in air distribution (e.g. pneumothorax, subcutaneous emphysema, pneumoperitoneum, etc).

- Differential Diagnosis Tree (A.1.2): Diagnoses, as opposed to radiographic findings, are those entities that radiologists can propose only as a possible list of differential diagnoses because the patient's clinical information and/or additional studies, such as other radiology tests or laboratory analyses, are required to interpret the radiographic pattern and suggest a diagnosis with certainty. The more a radiologist knows about the patient (clinical, exploration and laboratory data), the more accurate the diagnosis is. For example, an alveolar pattern is a radiographic finding that would prompt a very long list of differential diagnoses (including both infectious and non-infectious diseases such as lung edema, respiratory distress, etc.), but whenever it is present in a patient with fever, cough, leukocytosis, high CRP levels and crackles in the localization of the alteration, then it will almost certainly be pneumonia. This additional information is not always provided to the radiologist, who consequently lists the most probable differential diagnoses in the report. Up to 30% of false positive and 30% of false negative diagnoses of pneumonia based on the presentation and chest X-ray findings were reported by [31]. In this work, we did not access the patients' health record to confirm diagnoses, and those entities were, therefore, regarded as differential diagnoses.

- Topological Tree (A.1.3): This tree groups anatomical or spatial qualifiers in which radiographic findings and diagnoses are located. These spatial concepts include body parts, organs or organ components, chest anatomical regions, structures, and type of tissues. These concepts were first identified by means of regular expressions (see B) and then assigned to labels that were hierarchically organized in the topological tree.

2.4 Automatic labeling of the remaining reports

The reports that were not labeled in the previous stage (75%) were automatically tagged using a deep neural network classifier trained with the manually labeled data. We, therefore, used the manually annotated set of x-ray reports as training set. The goal was to produce a large dataset of annotated reports to be used as output of an image classifier trained with the x-rays.

Parameter	Value
Number of training sentences	20,439
Number of validation sentences	2,271
Vocabulary size	9,691
Average (min-max) number of tokens per sentence	7.1 (1-56)
Average (min-max) number of labels per sentence	1.33 (1-9)
Label space $|\mathcal{L}|$	193

Table 2: Descriptive statistics of the report sentences used to train the models. All sentences were manually annotated by trained physicians. Note that the reports consist of a sequence of sentences.

We define the task of annotating the medical entities as a multi-label text classification problem in which, for each sentence i, the goal is to predict $y_{i,\ell} \in 0, 1$ for all $\ell \in \mathcal{L}$, where $\mathcal{L}$ is the label space that includes the different radiographic findings (A.1.1) and differential diagnoses (A.1.2).

Our method does not employ any structured data or external information beyond the manually labeled data. The descriptive statistics of the training set and ground truth are summarized in Tab. 2.

In this method, each input instance corresponds to a sentence (denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, .., \boldsymbol{x}_N]$) composed of a sequence of pretrained word-embeddings $\boldsymbol{x}_i$ of dimension d_e for each of its N tokens, and its extracted labels encoded in a one-hot vector $\boldsymbol{y}$ of dimension $|\mathcal{L}|$. Word-embeddings of size $d_e = 100$ were trained on the full corpus of Spanish reports with FastText [32] in order to obtain the embeddings with the following parameters: Learning rate 0.025, size of the context window 5, number of epochs 55, minimum number of word occurrences 5, number of negative sampled 5, negative sampling loss function, sampling threshold 10^{-4}, length of char n-gram (min-max) $3 - 6$. We favored FastText over Word2Vec method so as to be able to encode unseen words that may potentially appear at inference time.

The four evaluated models are:

- A 1-Dimensional Convolutional Neural Network (CNN).
- A Recurrent Neural Network (RNN).
- A 1-Dimensional Convolutional Neural Network with attention mechanism (CNN-ATT).
- A Recurrent Neural Network with attention mechanism (RNN-ATT).

The CNN and RNN models were regarded as baseline models in order to compare them with the same architectures including attention mechanisms. The topologies of these systems are shown in Figs 6 and 7 respectively. All neural models were implemented using PyTorch [33] and the Ignite [3] library was used to help compact the programming code for common tasks such as training loop with metrics, early-stopping and model checkpointing.

2.4.1 CNN

In the baseline CNN, after inputing the text as a matrix of concatenated word-embeddings, the model computed a base representation of each sentence denoted as matrix $\boldsymbol{H} \in \mathbb{R}^{31x128}$ containing the horizontal concatenation of the n-grams spatial localizations with 128 features each. The sigmoid activations of a final linear layer outputted the $|\mathcal{L}|$ binary classification labels.

As shown in Fig. 6, each sentence in the CNN is represented as a matrix of $N = 56$ word-embeddings of $d_e = 100$ dimensions. The first convolutional layer contains 64 filters of dimension 3x1 with stride 1 and ReLu activation function, followed by a 2x1 max pooling layer with stride 1. There is a second convolutional layer containing 128 filters with the same dimensions, stride and activation function as the first layer. In this baseline CNN model, the attention module (highlighted in gray in Fig. 6) is replaced with a $|\mathcal{L}|$ fully connected layer with multi-label sigmoid activations.

2.4.2 RNN

In the baseline RNN, after inputing the text as a variable length sequence of word-embeddings, the model computed a base representation of each sentence, denoted as matrix $\boldsymbol{H} \in \mathbb{R}^{Nx256}$ containing the sequence of hidden states, in which each hidden state represents one token at one time step encoded with 256 features each. As occurred with the CNN, the sigmoid activations of a final lineal layer outputted the $|\mathcal{L}|$ binary classification labels.

As shown in Fig. 7,each sentence in the RNN is inputed at each time step as a sequence of variable length of N word-embeddings of $d_e = 100$ dimensions. There are two LSTM bi-directional layers containing 128 hidden units each.
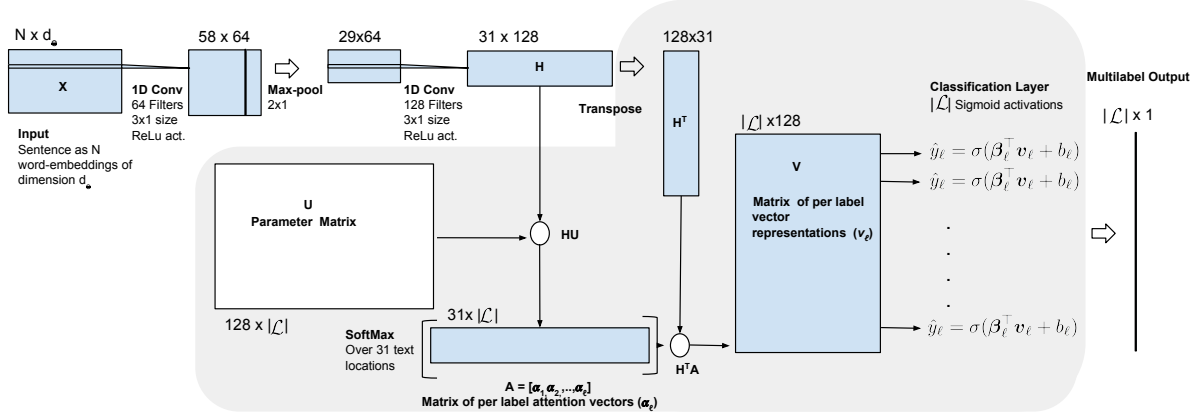
---

[3] https://pytorch.org/ignite/

Figure 6: Diagram of the CNN topology with attention mechanism highlighted in gray. Input is a representation of the text as a matrix of concatenated word-embeddings, and outputs are the most likely labels $\mathcal{L}$.
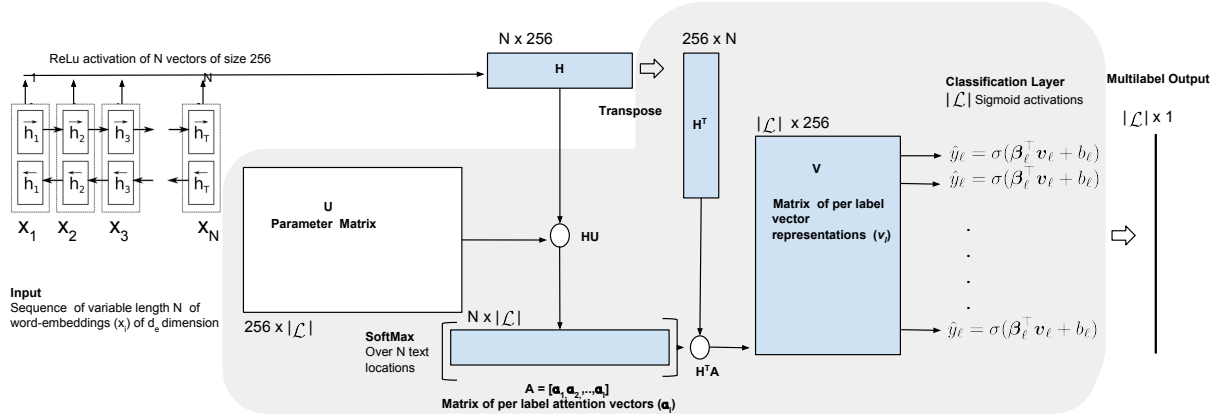

Figure 7: Diagram of the RNN topology with attention mechanism highlighted in gray. Input is a representation of the text as a sequence of variable-length word-embeddings, and outputs are the most likely labels $\mathcal{L}$.

In this baseline RNN model, the attention module (highlighted in gray in Fig. 7) is removed and the hidden state of the last time step of the RNN is used as the input for a final $|\mathcal{L}|$ fully connected layer with multi-label sigmoid activations.

2.4.3 Attention model

The attention mechanism, which is based on the work by [34], learns different text representations for each label. The underlying idea is that each snippet containing important information correlated with a label could be anywhere in the text and would differ among different labels.

In both the CNN-ATT and the RNN-ATT model, the final linear layer was replaced with an attention feed-forward module that first calculates attention vectors $\boldsymbol{\alpha}_\ell$ as a distribution of putative locations for each label ℓ in the sentence. As illustrated in Figs. 6 and 7, attention vectors are computed as:

$$\boldsymbol{\alpha}_\ell = \text{SoftMax}(\boldsymbol{H}\boldsymbol{u}_\ell) \tag{1}$$

where $\boldsymbol{u}_\ell$ is a vector parameter for label ℓ and $\text{SoftMax}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_{\forall i} \exp(\boldsymbol{x}_i)}$, where $\exp(\boldsymbol{x})$ is the element-wise exponentiation of the vector $\boldsymbol{x}$ as described in [34].

In a second step, we apply the attention vectors $\boldsymbol{\alpha}_\ell \in \boldsymbol{A}$ to the base representation of a sentence in order to compute a matrix of 128-dimensional weighted feature vectors $\boldsymbol{v}$ for each label, defined as $\boldsymbol{v}_\ell = \boldsymbol{H}^\top \boldsymbol{\alpha}_\ell$ where $\boldsymbol{v}_\ell \in \boldsymbol{V}$.

Finally, following the approach of [34], after applying a linear layer consisting of $|\mathcal{L}|$ vectors of weights $\boldsymbol{\beta}_l$ and a scalar bias b_ℓ, a sigmoid transformation is used to compute the probability $\hat{y}_\ell$ for each label.

$$\hat{y}_\ell = \sigma(\boldsymbol{\beta}_\ell^\top \boldsymbol{v}_\ell + b_\ell) \tag{2}$$

The training procedure minimized the binary cross entropy loss L in the case of all four topologies:

$$L_{BCE}(\boldsymbol{X}, \boldsymbol{y}) = -\sum_{\forall l}(y_\ell \log(\hat{y}_\ell) + (1 - y_\ell)\log(1 - \hat{y}_\ell)) \tag{3}$$

and the L2-norm of the model weights, using Adam optimizer [35] for the CNN and CNN-ATT models and RMSprop [36] for the RNN and RNN-ATT models.

3 Evaluation of automatic labeling

The following metrics were used to assess the performance of the automatic labeling method:

- Accuracy: The accuracy score provides the fraction of labels that were correctly detected. In this work, as we are dealing with multi-label classification, we obtain the average accuracy per sample: If all the predicted labels for an input sentence strictly match the true set of labels, then the accuracy for that sample is 1, and is otherwise 0.0. The average accuracy of all the samples in the test set is reported as the overall accuracy.

- MacroF1: Calculates metrics for each label, and finds their unweighted mean. This does not take label imbalance into account, as it places more emphasis on rare label predictions.

$$\text{MacroR} = \frac{1}{|\mathcal{L}|}\sum_{\ell=1}^{|\mathcal{L}|}\frac{\text{TP}_\ell}{\text{TP}_\ell + \text{FN}_\ell} \tag{4}$$

$$\text{MacroP} = \frac{1}{|\mathcal{L}|}\sum_{\ell=1}^{|\mathcal{L}|}\frac{\text{TP}_\ell}{\text{TP}_\ell + \text{FP}_\ell} \tag{5}$$

$$\text{MacroF1} = 2 \cdot \frac{\text{MacroR} \cdot \text{MacroP}}{\text{MacroR} + \text{MacroP}} \tag{6}$$

- MicroF1: Calculates metrics globally by counting the number of true positives, false negatives and false positives.

$$\text{MicroR} = \frac{\sum_{\ell=1}^{|\mathcal{L}|}\text{TP}_\ell}{\sum_{\ell=1}^{|\mathcal{L}|}(\text{TP}_\ell + \text{FN}_\ell)} \tag{7}$$

$$\text{MicroP} = \frac{\sum_{\ell=1}^{|\mathcal{L}|}\text{TP}_\ell}{\sum_{\ell=1}^{|\mathcal{L}|}(\text{TP}_\ell + \text{FP}_\ell)} \tag{8}$$

$$\text{MicroF1} = 2 \cdot \frac{\text{MicroR} \cdot \text{MicroP}}{\text{MicroR} + \text{MicroP}} \tag{9}$$

- WeightedF1: Calculates metrics for each label, and finds their average, weighted by support (the number of true instances for each label). This alters Macro to account for label imbalance and it may result in a score that is not between precision and recall.

The four models were trained and validated in the same random partition and early stopping was employed in order to maintain the best model leaving the max number of epochs up to 500. The hyper-parameters used to train the models are summarized in Tab 3. The accuracy metrics obtained in the multi-label annotation of radiographic findings and differential diagnoses on text are summarized in Tab. 4.

Cross-validation experiments using k-folds with $k = 11$ were performed in order to attain accuracy curves and compare the learning pattern up to 150 epochs for the four models (CNN, CNN-ATT, RNN and RNN-ATT). The results are

Parameter	Values	CNN	RNN
Batch size	$256, 512, 1024, 2048$	1024	1024
L2 Penalty	$0, 10^{-3}, 10^{-2}$	0	0
Learning rate	$10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 10^{-2}$	10^{-4}	10^{-2}
Filter size	$[2, 10]$	3	
Number of filters second conv layer	$64, 128, 256$	128	
Dropout probability	$[0.1, 0.8]$	0.4	0.4
Number of hidden LSTM units	$64, 128, 256, 512$		128
Number of hidden LSTM layers	$[1, 3]$		2

Table 3: Hyperparameter ranges evaluated and best values chosen for the CNN and RNN based models selected by grid-search and manual fine tuning using MicroF1 as performance measure in the validation set.

Model	Epochs	Accuracy	MacroF1	MicroF1	WeightedF1
Validation Set					
CNN	94	0.706	0.384	0.837	0.814
CNN-ATT	230	0.806	0.458	0.902	0.886
RNN	78	0.853	0.483	0.913	0.903
RNN-ATT	41	**0.864**	**0.491**	**0.924**	**0.918**
Test Set					
RNN-ATT	41	0.857	0.491	0.939	0.926

Table 4: MicroF1 of all the models in a validation set of 2,271 sentences with a training set of 20,439 sentences. Best results are shown in bold type. The best model from among the four different architectures was further tested in an independent random sample of 500 sentences drawn from the PadChest Dataset and manually labeled.

shown in Fig. 8. These experiments allow us to understand the dispersion in accuracy attributable to the differences in the data distribution from the training and validation sets.

We draw the following conclusions:

- The RNN outperformed the CNN.

- Attention mechanisms increased the performance of both the CNN and the RNN models with the following particularities: The gain in the overall MicroF1 was more pronounced in the CNN-ATT model, but it had a slower but steadier increase when compared to the plain CNN. Conversely, the accuracy gain of the RNN-ATT versus the plain RNN was less appreciable but the model learned faster in early epochs.

- Accuracy suffers from a higher variance in the CNN-ATT as can be seen from the wider standard deviation interval compared to that of the RNN-ATT.

- For all four methods, in an attempt to reduce the variability in validation metrics, we experimented reducing the size of the training set in favor of a larger validation set but observed that the model's predictive performance deteriorated as regards both the training and validation metrics, particularly in the CNN architecture. We consequently, hypothesize that this architecture seems to benefit more from adding a larger number of instances, but this could not be proven given the available size of the labeled dataset.

A random sample of 500 sentences drawn from the partition of the PadChest dataset and not used to train or validate the model was further labeled manually and used as an independent test set. In order to test the suitability of the trained model as regards its generalization to the whole PadChest dataset, this sample test was drawn from x-reports belonging to a different period (2009 to 2013) with respect to those of the training and validation sets (2014-2017).

The best model from among the four different architectures was further tested in this test set, obtaining a MicroF1 score of 0.93, as shown in Tab. 4. This result was similar to those achieved in the validation set, illustrating its suitability for use in the automatic annotation of the PadChest dataset samples that were not labeled manually. The RNN-ATT model was consequently used to extract the annotations for the remaining PadChest dataset, as shown in Tab. 7.

4 Dataset overview

The dataset consists of 160,868 labeled chest x-ray images from 69,882 patients, acquired in a single institution between 2009 and 2017 (see Tab. 7). The patients had a mean of 1.62 chest x-ray studies performed at different time points (from 1 to 119). Each study contains one or more images corresponding to different position views, mainly P-A and lateral, and is associated with a single radiography report describing the results of all position views in a common text
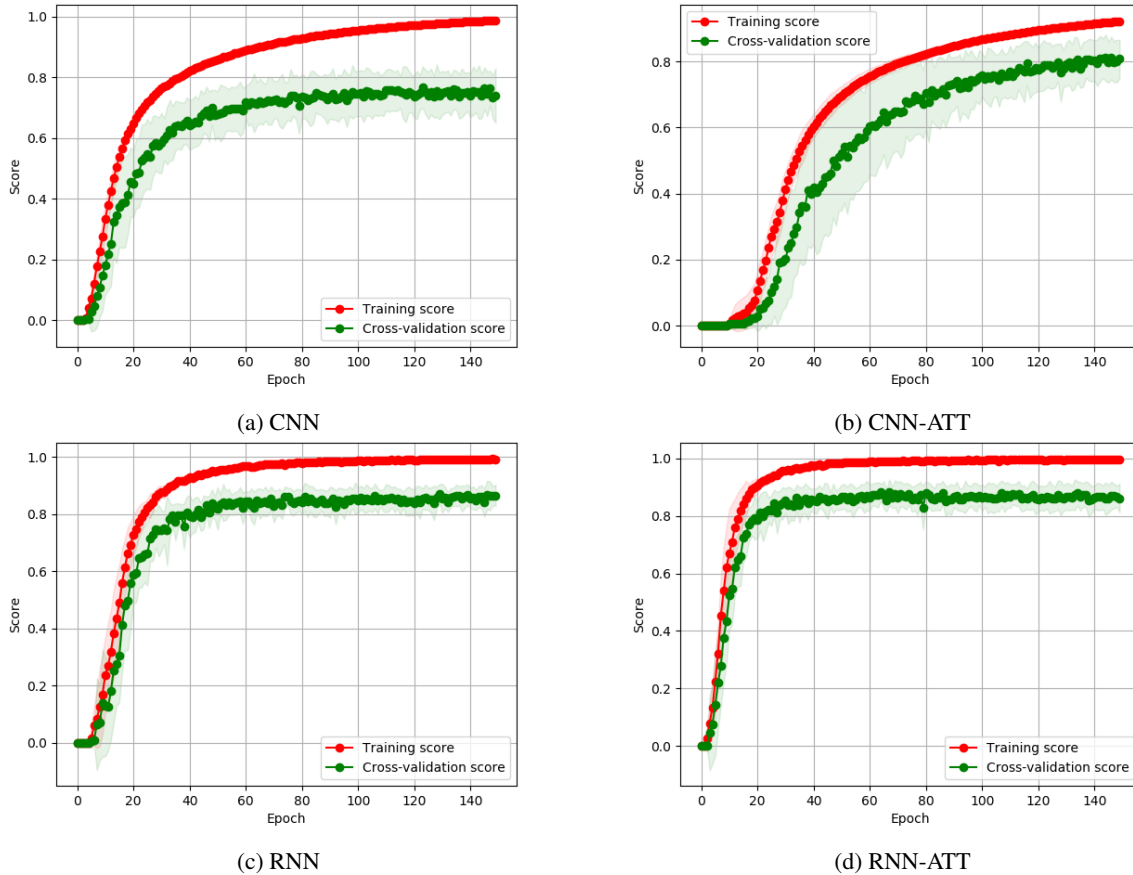
Figure 8: Learning curves: Each curve shows the average and standard deviation (shaded regions) of the MicroF1 precision score in 150 epochs for each model after cross-validation. The training and validation set contained 20,439 and 2,271 samples respectively.

section. The mean of the chest-x ray images was 2.37 per patient. The dataset generated provides two types of fields for each chest-x ray image: those fields with the suffix DICOM 6 contain the values of the original field in the DICOM standard and the remaining fields 5 enrich the PadChest dataset with additional processed information.

4.1 Patient Characteristics

The patients' ages range from 0 to 105 years, with a mean of 58.5 and a median of 62. The distribution of number of images by age (see Fig. 9) is skewed toward older ages with a long tail for ages under 40, and there are also differences according to gender in the age interval from 43 to 60 years, in which more x-rays correspond to women, contrasting with the age range from 71 to 75 in which more correspond to men. In the final labeled dataset, 80,923 images correspond to women and 79,923 to men. The median birth year of the population was 1953, with a standard deviation of 20 within a range from 1904 to 2017, as shown in Tab. 9.

As shown in Fig. 10, the radiation exposure (mAs), which is a derived field calculated from the x-ray tube current (mean 250) and the exposure time (mean 31, median 7, max 1,000), largely varied as regards modality (higher in computer CX than in digital radiographies DX), position view (higher in lateral than in frontal views) and year, and had a decreasing trend over time.

4.2 Chest X-ray Diagnoses, Radiographic Findings and Anatomical Locations

There are 68,855 studies (17,513 manually annotated by physicians) with radiographic findings and/or differential diagnoses, and 37,871 normal studies (of which 9,404 are manually labeled). Of these studies, 1.8% were reported to have sub-optimal technique and 1.2% (0.6% manual) were excluded because no information could be extracted from the reports, as shown in Fig. 11.

15

(a) Ages by gender

(b) Birth year

(c) Images per patient

(d) Year of study

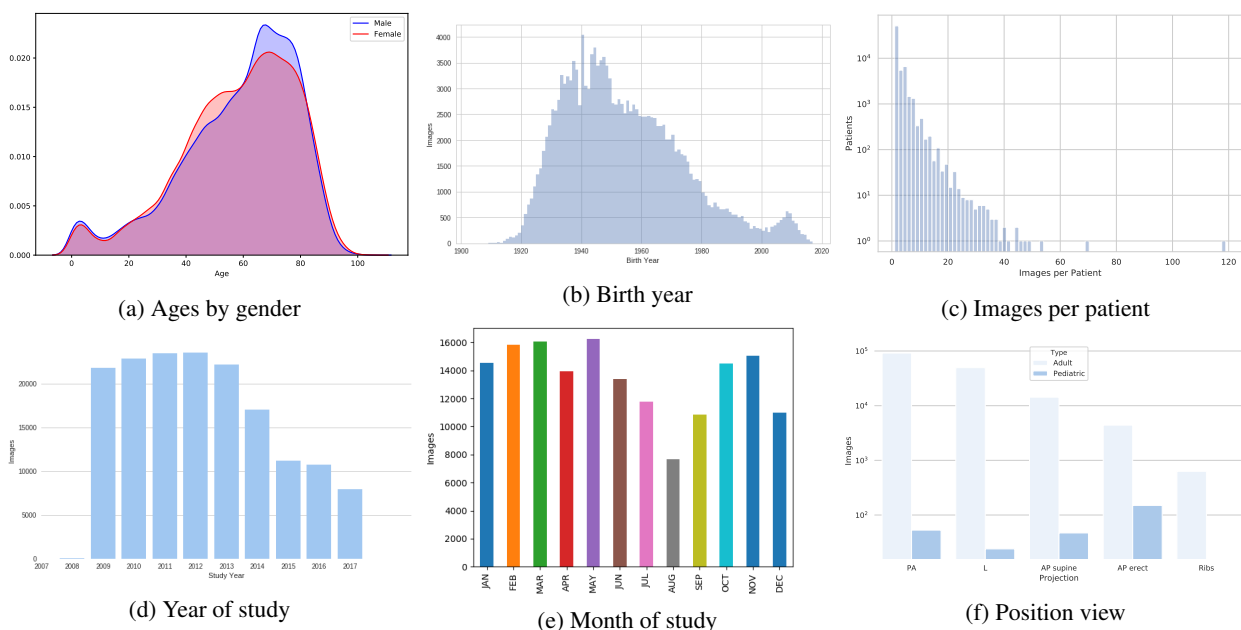(e) Month of study

(f) Position view

Figure 9: Dataset descriptive statistics. a) Kernel density estimation of the probability distribution of images by patient's age and gender. b) Statistics of the year of birth. c) Number of different images per patient. d), e) and f) Image year, month and type of projection: Postero-anterior (PA), Lateral (L), Antero-posterior (AP) supine and erect histograms, respectively. Note that in the case of the subset of images with projection not retrievable (UNK) from DICOM, a CNN model was used to classify them into L and PA.

(a) Radiation exposure by projection

(b) Radiation exposure by year

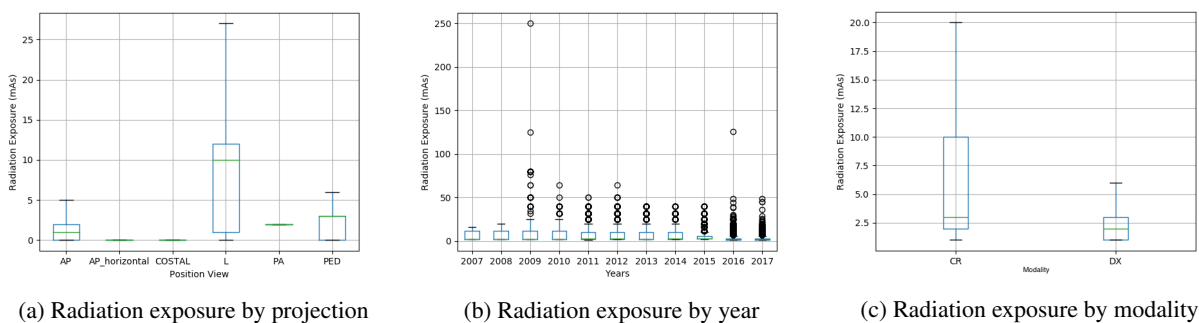(c) Radiation exposure by modality

Figure 10: Dataset radiation statistics. Radiation exposure was plotted against a) projections, b) years of study and c) modality (computed radiography CR versus digital radiography DX).

Name	Description
ImageID	Image identifier
ImageDir	Zip folder containing the image
StudyID	Study identifier
PatientID	Patient's code
PatientBirth	Year in format `YYYY`
Projection	Classification of the 5 main x-ray projections: `PA` (Postero-Anterior standard), `L` (Lateral), `AP` (Antero-Posterior erect or vertical), `AP-horizontal` (Antero-Posterior horizontal), `COSTAL` (rib views)
Pediatric	`PED` if the image acquisition followed a pediatric protocol
MethodProjection	The method applied in order to assign a projection type: based on a manual review of DICOM fields or based on the classification output of a pre-trained ResNet50 model for images without DICOM information about the projection.
ReportID	Integer identifier of the report
Report	A text snippet extracted from the original report containing the radiographical interpretation. The text is preprocessed, while the words are stemmed and tokenized. Each sentence is separated by '.'
MethodLabel	The method applied for manual labeling, manually by physicians (`Physician`) or supervised (`RNN_model`)
Labels	A sequence of unique labels extracted from each report
Localizations	A sequence of unique anatomical locations extracted from each report. Each anatomical location is always preceded by the token `loc`
LabelsLocalizationsBySentence	Sequences of labels followed by its anatomical locations. Each single sequence corresponds to the labels and anatomical locations extracted from a sentence and repeats the pattern formed of one label followed by no or many locations for this label `[ label, (0..n) loc name ]`. The sequences are ordered by sentence order in the report
LabelCUIS	A sequence of UMLS Metathesaurus CUIs corresponding to the extracted labels in the Labels field
LocalizationsCUIS	A sequence of UMLS Metathesaurus CUIs corresponding to the extracted anatomical locations in the Localizations field

Table 5: Dataset Fields: All additional processed fields that are different from original DICOM fields. Additional information on UMLS Metathesaurus CUIs can be found at `https://uts.nlm.nih.gov/home.html`.

Differential diagnoses were annotated with 19 different labels. In the 27,726 differential diagnoses reported (5,472 manual), the top 4 were COPD (14,557 studies, 3,017 manual), pneumonia (5,934, 1,158 manual) followed by heart insufficiency and pulmonary emphysema, as shown in Fig. 12.

Differential diagnoses were annotated with 19 different labels. In the 27,726 differential diagnoses reported (5,472 manual), the top 4 were COPD (14,557 studies, 3,017 manual), pneumonia (5,934, 1,158 manual) followed by heart insufficiency and pulmonary emphysema as shown in Fig. 12.

Radiographic findings were annotated with 170 different labels. As shown in Fig. 13, of the 177,628 (45,375 manual) findings reported, the top 10 were lung infiltrates (19,850), with an interstitial pattern in 7,001 and an alveolar pattern in 7,637 studies, followed by thoracic cage deformations (mainly owing to scoliosis), aortic elongations or dilations (10,154), pulmonary chronic changes, cardiomegaly, surgery marks, vertebral degenerative changes, atelectasis, pleural effusion (8,593) and hilar enlargements (4,895). Masses, nodules, and pseudonodules were reported in 1,026, 2,669 and 2,519 studies, respectively. Aortic atheromatosis, as detected in x-rays by means of vascular calcifications, were reported in 1,813 (1,125 manual) studies and bone fractures in 3,367 (820 manual). All artifacts, such as electric devices (4,286 with 1,063 manual), tubes (10,776 with 945 manual) and catheters (6,167 with 654 manual), were also labeled in order to include their different types. With regard to rare conditions, the Chilaiditi sign was reported in only 14 (6 manual) studies and dextrocardia was present in 14 (4 manual) studies (0.014%).

Anatomic locations were annotated with 104 different labels. In the 249,469 locations specified, the majority of the findings and/or diagnoses were localized in the lung fields (78,888 findings, the majority being - 34,661 - in the lower lung fields), followed by mediastinum (39,134 findings, the majority being in the lower mediastinum in the cardiac region), bone (12,862 findings) and soft tissue (3,567 findings). There were 13,935 findings reported in the pleural region, 11,349 in the hilar region, 9850 aortic findings and 3,777 diaphragmatic findings, as shown in Fig. 15.

4.3 Image projections and additional characteristics

The mean and standard deviation images were obtained for visualization purposes. They were calculated from a random sample of 500 training examples.

These images showed relevant differences among the different x-ray projections (see Fig. 17). In PA, the mean shows that images were fairly aligned to the center, while the standard deviation for the same random sample of 500 images,
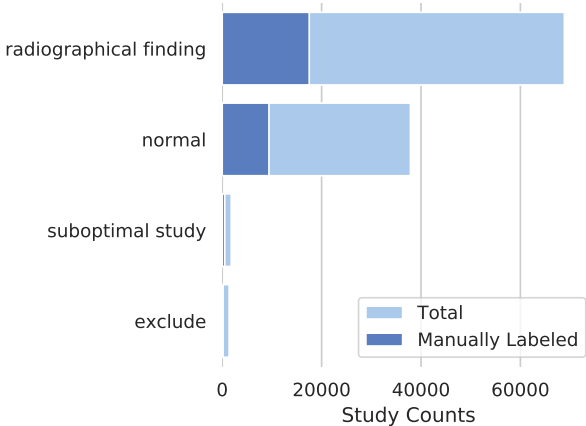
Figure 11: Chest x-ray report types. Data are shown for both the manual (dark color) and automatically labeled dataset (light color).
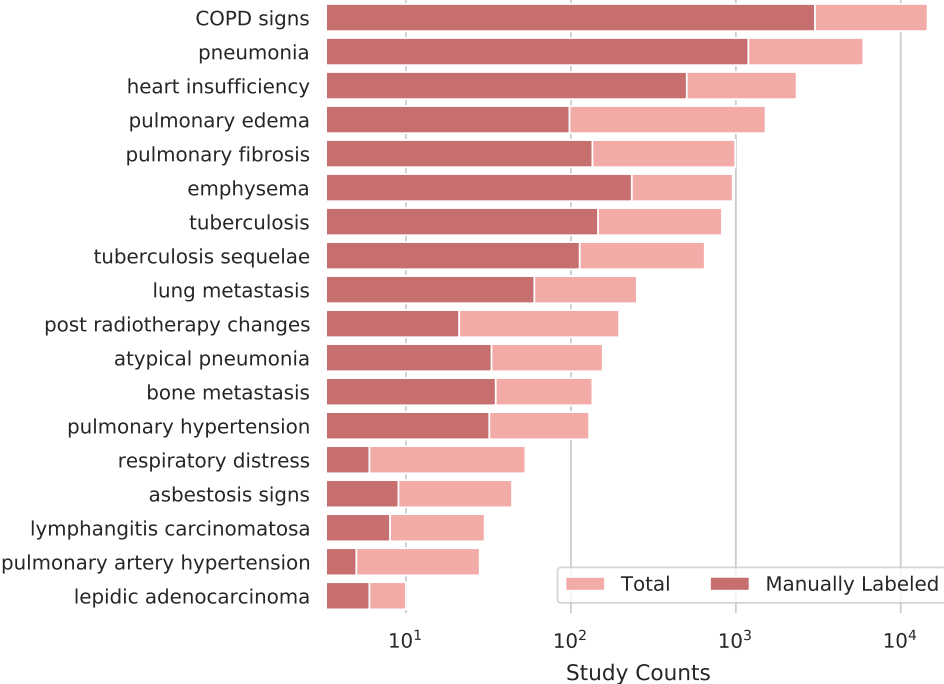

Figure 12: Most frequent differential diagnoses. Data are shown for both the manual (dark color) and automatically labeled dataset (light color). See A.1.2 for counts of labels on each hierarchical tree.

Name	DICOM tag	Description
StudyDate_DICOM	0008,0020	YYYY-MM-DD Starting date of the study
PatientSex_DICOM	0010,0040	Sex of the patient. M male, F female, O other
ViewPosition_DICOM	0018,5101	Radiographic view of the image relative to the subject's orientation
Modality_DICOM	0008,0060	Type of equipment used to acquire the image data
Manufacturer_DICOM	0008,0070	Manufacturer of the equipment that produced the images
PhotometricInterpretation_DICOM	0028,0004	Intended interpretation of the pixel data represented as a single monochrome image plane. MONOCHROME1: the minimum sample value should be displayed in white. MONOCHROME2: the minimum sample value is intended to be displayed in black
PixelRepresentation_DICOM	0028,0103	Data representation of the pixel samples, unsigned integer or 2's complement
PixelAspectRatio_DICOM	0028,0034	Ratio of the vertical size and horizontal size of the pixels in the image specified by a pair of integer values, in which the first value is the vertical pixel size and the second value is the horizontal pixel size
SpatialResolution_DICOM	0018,1050	The inherent limiting resolution in mm of the acquisition equipment
BitsStored_DICOM	0028,0101	Number of bits stored for each pixel sample
WindowCenter_DICOM	0028,1051	Window width for display
WindowWidth_DICOM	0028,1050	Window center for display.
Rows_DICOM	0028,0010	Number of rows in the image
Columns_DICOM	0028,0011	Number of columns in the image
XRayTubeCurrent_DICOM	0018,1151	X-ray Tube Current in mA
ExposureTime_DICOM	0018,1150	Duration of x-ray exposure in msec
Exposure_DICOM	0018,1152	Exposure expressed in mAs, calculated from Exposure Time and x-ray Tube Current
ExposureInuAs_DICOM	0018,1153	Exposure in μAs
RelativeXRayExposure_DICOM	0018,1405	Relative x-ray exposure on the plate

Table 6: DICOM dataset fields: DICOM® (Digital Imaging and Communications in Medicine) is the international standard employed to transmit, store, retrieve, print, process, and display medical imaging information. Detailed descriptions of DICOM standard fields can be found in [37].

	Patients	Reports	Images
Initial Set	69,882	115,678	168,171
Excluded Set	4,744	5,864	7,303
Photometric interpretation			5,652
Modality			871
Study Report			705
Position View			42
Protocol			23
Pixel Data			10
Ground Truth			
Manual Report Labeling	24,491	27,593	39,039
Automatic Report Labeling	43,134	82,338	121,829
Final Labeled Set	67,625	109,931	160,868

Table 7: PadChest global statistics.

in which higher values are whiter, indicates a higher variation at the boundaries when compared to the center. In AP supine position, the mean heart silhouette was enlarged. Vertical and horizontal lines in the pediatrics standard deviation illustrate that there are many sizes of frames enclosing the chest and part of the abdomen.

5   Conclusions

In this work, we present one of the largest public chest x-ray dataset, which contains more than 160K images and Spanish reports labeled with 299 different medical entities mapped on to UMLS CUIs. It is expected that this dataset and the projects derived from it will ultimately assist in the development of diagnostic decision support systems, thus increasing clinical practice efficiency. The main contributions of PadChest are the following:

- This is the first large-scale exploitable dataset available in Spanish. Given that it has been labeled with standardized medical codes, it is ready for further image exploitation, regardless of the language required.

- This is one of the largest publicly available labeled dataset that, in contrast to previously published works, includes different projection views and additional metadata, which is essential information if the models are to be trained adequately, as discussed previously.

	Images
Projection and Positioning	
PA	96,010
L	51,124
AP-Horizontal	14,355
AP-Vertical	5,158
Costal	631
Pediatric	274
Gender	
Male	80,923
Female	79,923

Table 8: Descriptive statistics: Categorical variables.

	Mean ± Std	Min	Max
Year of Birth	1953 ± 20	1904	2017
Age	58.5 ± 20	0	105
Studies per patient	1.62 ± 1.66	1	116
Images per patient	2.37 ± 2.34	1	119
Images per study	1.46 ± 0.53	1	11
Rows per image	2254 ± 739	118	4280
Columns per image	2299 ± 748	20	4280
Spatial Resolution	0.16 ± 0.02	0.1	0.2

Table 9: Descriptive statistics: Continuous variables.

- Trained physicians manually annotated 27% of the dataset samples, and a RNN-ATT method was used to label the remaining reports.

- We propose hierarchical taxonomies in order to categorize radiographic findings, differential diagnoses and anatomical locations following UMLS CUI standard codes. To the best of our knowledge, available labeled datasets have a lower number of different labels.

The manually labeled dataset was used to train models for the automatic annotation of a Spanish corpus of x-ray reports. The best model was used to label 73% of the PadChest dataset and achieved a 0.93 MicroF1 score for an independent test set, proving the robustness of the proposed approach as regards labeling the full dataset.

One downside is that all the data were drawn from a single institution, but the fact that the training and test sets belong to different year intervals (2014-2017 and 2009-2014, respectively), thus increasing the variability of the reporting radiologists, supports that the model has the potential to be generalized. This would,however, have to be confirmed by using reports from other Spanish institutions.

This dataset can be used to train models to predict thoracic pathologies affecting lung, cardiovascular and bone structures from x-ray images.

We are aware that the dataset has some limitations. First limitation is the information bias inherent to employing the reports as the only source of the ground-truth. Specifically, the image could have radiographic abnormalities which are not described in the reports. For example, the presence of medical devices, such as pacemakers, catheters or prosthesis, or the presence of chronic and/or non severe conditions, such as chronic bone degenerative changes, granulomas, prior surgery marks, etc., are not systematically reported in clinical practice by radiologists and the decision to include them on the report depends on the clinical context, the requesting department and even on the radiologists' work-load. Although these entities are not usually clinically relevant, we decided to label them so as to allow the models to assess their influence as potential bias or confounding factors when predicting clinically relevant entities. Moreover, the retrieval of images with these radiographic findings may be useful to help build more focused datasets with, for example, the goal of identifying misplaced medical devices. This would be useful when employing a chest x-ray to verify the correct position of the distal end after having inserted a central catheter. Although additional labeling would be required for this purpose, this dataset would initially help retrieve the relevant images.

Second, the severity of medical entities when described is not currently captured in the labels. The adjectives usually employed to describe severity are: severe, intense, important, marked, light, faint, thin, slight, minimal, prominent, discreet, moderate, etc. Nonetheless, anatomic extensions such as bi-lateralization, multi-lobality, etc., which usually correlate with severity, are included in the topological labels.
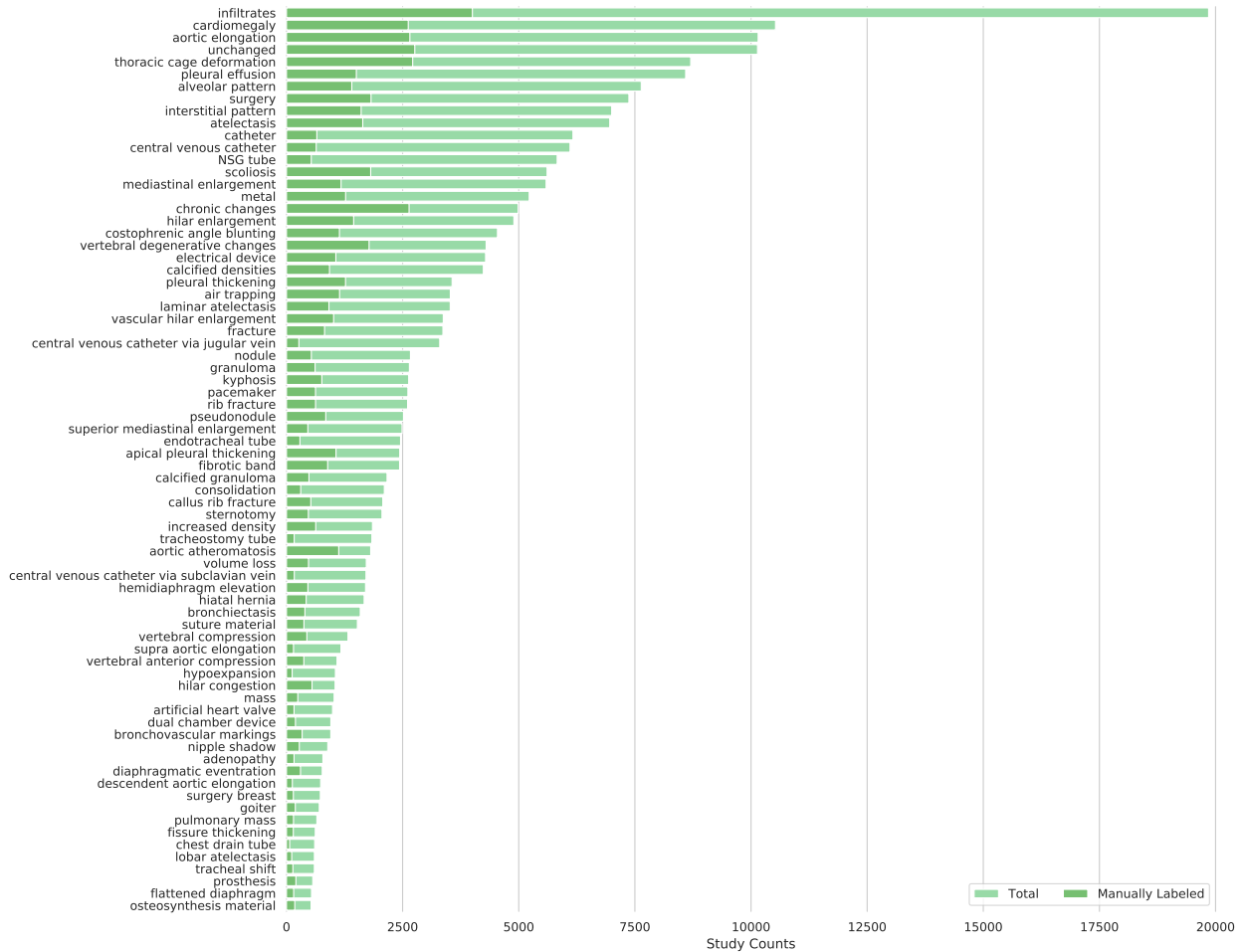
Figure 13: Most reported radiographic findings.

Figure 14: Most frequent radiographic findings. Labels are shown for both the physician (dark color) and automatically labeled dataset (light color). See A.1.1 for counts of labels on each hierarchical tree.

Third limitation is the selection bias introduced by including only those x-ray studies with retrievable reports from PACs. Radiology explorations prescribed to answer questions of type yes/no and not for diagnosis are not necessarily accompanied by reports. This selection bias may be, along with other recognized factors such as differences as regards disease burden, accountable for the differences in the pathology frequencies in data-sets from different institutions.

One important warning is that datasets across different institutions may vary greatly, and this has been proved to greatly influence the generalization of Machine Learning Decision Support Systems (ML-DSS), particularly if confounding variables exist in internal data that do not exist in external data. A recent cross-sectional study [9] showed that CNN models trained to predict pneumonia from x-rays learned to exploit large differences in pathology prevalence between hospital systems in training data by calibrating their predictions to the baseline prevalence in each hospital system rather than exclusively discriminating on the basis of direct pathology findings. In addition, the CNNs robustly identified a hospital system and a department within a hospital, which had large differences as regards disease burden and may have confounded the predictions. The authors caution that the estimates of CNN performance based on test data from hospital systems used for model training may overstate their real-world performance.

Before ML-DSS models can be confidently deployed in real-world clinical settings, two important requirements need to be addressed 1) controlling for confounding factors, this being particularly true in the case of DL models in which features are learned by the model rather than being engineered and predefined by researchers, and 2) prospectively evaluating the systems in multi-centric comparative studies that define target populations and adequate primary and secondary endpoints measuring health outcomes and comparing them with usual practice.

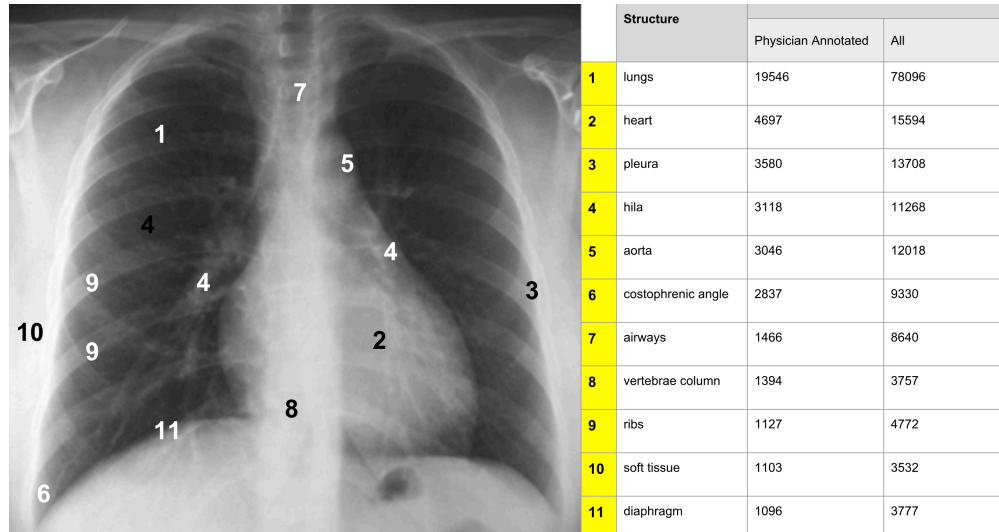| Structure		Physician Annotated	All
1	lungs	19546	78096
2	heart	4697	15594
3	pleura	3580	13708
4	hila	3118	11268
5	aorta	3046	12018
6	costophrenic angle	2837	9330
7	airways	1466	8640
8	vertebrae column	1394	3757
9	ribs	1127	4772
10	soft tissue	1103	3532
11	diaphragm	1096	3777

Figure 15: Most common locations of radiographic findings and differential diagnoses grouped by anatomical regions. See A.1.3 for counts on the locations tree.

Sound study designs would make it possible to consistently evaluate the performance of the systems, including their capacity to generalize across not only departments and hospitals, but also different populations and diversities of clinical settings.

Finally we provide the following advice for any researchers who intend to use PadChest for the task of predicting pathologies from x-rays:

1. To explore the use the PadChest dataset in conjunction with x-ray datasets from other institutions to help increase generalization.

2. To explore multi-input models and use all provided context information -including details of image acquisition, manufacturer, projection and clinical labels- to help increase the understanding of confounding factors exploited by ML-DSS models and to develop methods to control them, like for instance adjusting by expected prevalence of diseases.

3. To acknowledge the different nature of radiographic findings versus differential diagnoses, which has been commonly overlooked in prior works (for instance, the task of predicting cardiomegaly has been regarded as same type of problem as predicting pneumonia) and to explore different methods and evaluation criteria for each of them.

4. To explore models designed for dealing with labels having hierarchical relationships.

5. To acknowledge the information bias inherent to employing the reports as the only source of the ground-truth.

Acknowledgement

Figure 16: Most common locations of radiographic findings and differential diagnoses. See A.1.3 for counts on the locations tree.

(a) P-A   (b) Lateral   (c) AP supine   (d) AP erect   (e) Ribs   (f) Pediatric

(g) P-A   (h) Lateral   (i) AP supine   (j) AP erect   (k) Ribs   (l) Pediatric
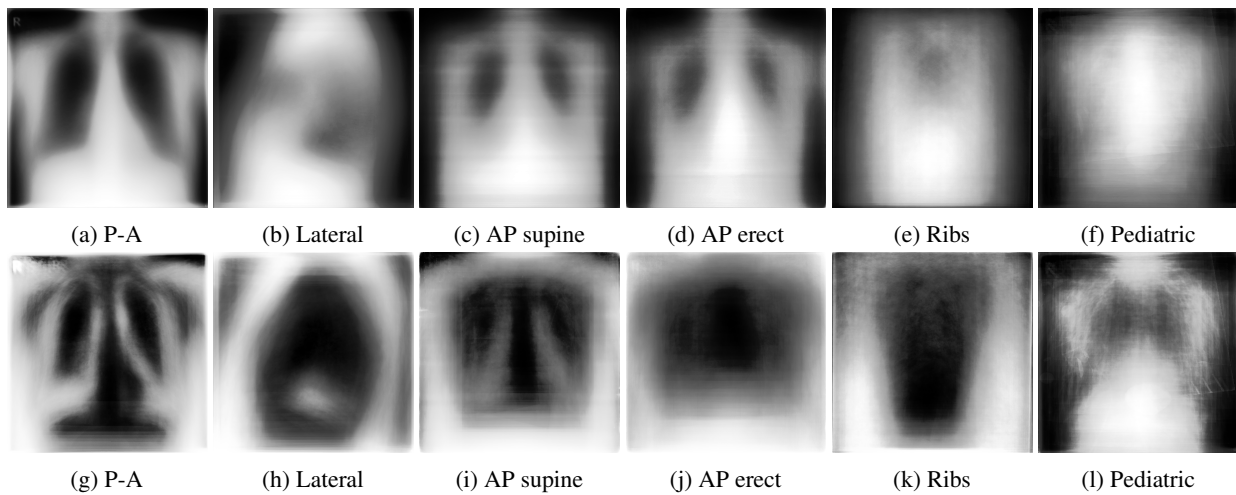
Figure 17: From a) to f): Mean image by position view. From g) to l): Standard deviation image by position view.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[3] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical engineering online*, 17(1):113, 2018.

[4] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017.

[5] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[6] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058, 2018.

[8] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[9] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018.

[10] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

[11] Hoo-Chang Shin, Le Lu, and Ronald Summers. *Natural Language Processing for Large-Scale Medical Image Analysis Using Deep Learning*. 12 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[13] Chunhua Weng, Samson W. Tu, Ida Sim, and Rachel Richesson. Formal representation of eligibility criteria: A literature review. *Journal of Biomedical Informatics*, 43(3):451 – 467, 2010.

[14] Aurelia Bustos and Antonio Pertusa. Learning eligibility in cancer clinical trials using deep neural networks. *Applied Sciences*, 8(7):1206, 2018.

[15] Sungweon Ryoo and Hee Jin Kim. Activities of the korean institute of tuberculosis. *Osong public health and research perspectives*, 5:S43–S49, 2014.

[16] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

[17] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[18] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

[19] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[20] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *The Journal of Machine Learning Research*, 17(1):3729–3759, 2016.

[21] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[22] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37, 2015.

[23] Jose Salinas Serrano, Maria de la Iglesia Vaya, Luis Marti-Bonmati, Rosa Valenzuela, and Miguel Cazorla. R&d cloud ceib: Management system and knowledge extraction for bioimaging in the cloud. *Distributed Computing and Artificial Intelligence*, 151:331–338, 01 2012.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Edward Loper and Steven Bird. *NLTK: The Natural Language Toolkit*. 2002.

[26] A Cichocki and A H Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, E92-A(3):708–721, 2009.

[27] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[28] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[30] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[31] Yann-Erick Claessens, Marie-Pierre Debray, Florence Tubach, Anne-Laure Brun, Blandine Rammaert, Pierre Hausfater, Jean-Marc Naccache, Patrick Ray, Christophe Choquet, Marie-France Carette, et al. Early chest computed tomography scan to assist diagnosis and guide treatment decision for suspected community-acquired pneumonia. *American journal of respiratory and critical care medicine*, 192(8):974–982, 2015.

[32] Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 3–7, 2017.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. *Automatic differentiation in PyTorch*. 2017.

[34] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1101–1111, 2018.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Tijmen Tieleman and Geoffery Hinton. Rmsprop gradient optimization, 2014. url-http://www.cs.toronto.edu/tijmen/csc321/slides/lecture_slides_lec6. pdf.

[37] Oleg S. Pianykh. *Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide*. Springer Science & Business Media, 2009.

A   Hierarchies of labels

A.1   Counts of annotated labels in x-ray studies

27,593 study reports were manually labeled and reviewed by trained physicians. Medical entities were extracted as per physician criteria and mapped to NLM Unified Medical Language System (UMLS) controlled biomedical vocabulary unique identifiers referred as CUIs. Medical entities whose exact meaning could not be found in any term of the UML metathesaurus but they were relevant for annotation purposes were also extracted using the physician's label but a CUI was not assigned. Four additional labels with special coding rules were explicitly defined: normal, exclude, unchanged and suboptimal study.

A.1.1   Hierarchy of labels of radiographic findings

A total of 177,628 radiographic findings (being 45,375 of those manually annotated in 27,593 study reports) were annotated in all the study reports. The labels "normal", "exclude" and "suboptimal study" are counted separately.

label [CUI, study counts, study counts with child labels]
normal [CUI:C0205307, counts:37871]
exclude [counts:1392]
suboptimal study [CUI:C2828075, counts:1813]
radiological finding [CUI:C0436485]
├── unchanged [counts:10140, 10140]
├── obesity [CUI:C0028754, counts:30, 30]
├── chronic changes [CUI:C0742362, counts:4988, 4988]
├── calcified densities [CUI:C2203586, counts:849, 4238]
│   ├── calcified granuloma [CUI:C0333404, counts:2165, 2165]
│   ├── calcified adenopathy [counts:434, 442]
│   │   └── calcified mediastinal adenopathy [counts:8, 8]
│   ├── calcified pleural thickening [counts:450, 450]
│   ├── calcified pleural plaques [counts:160, 160]
│   ├── heart valve calcified [CUI:C2073448, counts:129, 129]
│   └── calcified fibroadenoma [counts:43, 43]
├── granuloma [CUI:C0235557, counts:481, 2646]
│   └── calcified granuloma [CUI:C0333404, counts:2165, 2165]
├── end on vessel [counts:152, 152]
├── adenopathy [CUI:C0478664, counts:350, 784]
│   └── calcified adenopathy [counts:434, 434]
├── nodule [CUI:C0034079, counts:2420, 2669]
│   └── multiple nodules [CUI:C2073563, counts:249, 249]
├── pseudonodule [counts:1478, 2519]
│   ├── nipple shadow [counts:889, 889]
│   └── end on vessel [counts:152, 152]
├── abscess [CUI:C0024110, counts:9, 9]
├── cyst [CUI:C2073485, counts:5, 5]
├── cavitation [CUI:C0578537, counts:367, 367]
├── fibrotic band [CUI:C0865843, counts:2435, 2435]
├── volume loss [CUI:C3203358, counts:1718, 1718]
├── hypoexpansion [counts:1053, 1053]
├── bullas [CUI:C0241982, counts:418, 418]
├── pneumothorax [CUI:C2073565, counts:411, 440]
│   └── hydropneumothorax [CUI:C0020303, counts:29, 29]
├── pneumoperitoneo [CUI:C0032320, counts:56, 56]
├── pneumomediastinum [CUI:C2073636, counts:18, 18]
├── subcutaneous emphysema [CUI:C0038536, counts:190, 190]
├── hyperinflated lung [CUI:C0546312, counts:409, 409]
├── flattened diaphragm [CUI:C2073504, counts:539, 539]
├── lung vascular paucity [counts:108, 108]
├── air trapping [CUI:C0231819, counts:3527, 3527]
├── bronchiectasis [CUI:C0006267, counts:1588, 1588]

```
├── infiltrates [CUI:C0277877, counts:5212, 19850]
│   │   ├── interstitial pattern [CUI:C2073538, counts:6060, 7001]
│   │   │   ├── ground glass pattern [CUI:C3544344, counts:213, 213]
│   │   │   ├── reticular interstitial pattern [counts:427, 427]
│   │   │   ├── reticulonodular interstitial pattern [CUI:C2073672, counts:175, 175]
│   │   │   └── miliary opacities [CUI:C2073583, counts:126, 126]
│   │   └── alveolar pattern [CUI:C1332240, counts:5384, 7637]
│   │       ├── consolidation [CUI:C0521530, counts:1959, 2106]
│   │       │   └── air bronchogram [CUI:C3669021, counts:147, 147]
│   │       └── air bronchogram [CUI:C3669021, counts:147, 147]
├── bronchovascular markings [CUI:C2073518, counts:954, 954]
├── air fluid level [CUI:C0740844, counts:115, 115]
├── increased density [CUI:C1443940, counts:1851, 1851]
├── atelectasis [CUI:C0004144, counts:2666, 6959]
│   │   ├── total atelectasis [CUI:C0264497, counts:24, 24]
│   │   ├── lobar atelectasis [counts:601, 601]
│   │   ├── segmental atelectasis [counts:104, 104]
│   │   ├── laminar atelectasis [counts:3525, 3525]
│   │   ├── round atelectasis [CUI:C2062952, counts:3, 3]
│   │   └── atelectasis basal [CUI:C0746053, counts:36, 36]
├── mediastinal shift [CUI:C0264576, counts:39, 39]
├── azygos lobe [CUI:C0265794, counts:327, 327]
├── fissure thickening [counts:162, 620]
│   │   ├── minor fissure thickening [counts:342, 342]
│   │   ├── major fissure thickening [counts:30, 30]
│   │   └── loculated fissural effusion [counts:86, 86]
├── pleural thickening [CUI:C0264545, counts:677, 3565]
│   │   ├── apical pleural thickening [counts:2438, 2438]
│   │   └── calcified pleural thickening [counts:450, 450]
├── pleural plaques [CUI:C0340030, counts:15, 175]
│   │   └── calcified pleural plaques [counts:160, 160]
├── pleural effusion [CUI:C2073625, counts:8289, 8593]
│   │   ├── loculated pleural effusion [CUI:C0747639, counts:185, 185]
│   │   ├── loculated fissural effusion [counts:86, 86]
│   │   ├── hydropneumothorax [CUI:C0020303, counts:29, 29]
│   │   ├── empyema [CUI:C0014009, counts:4, 4]
│   │   └── hemothorax [CUI:C0019123, counts:0, 0]
├── pleural mass [CUI:C1709576, counts:2, 2]
├── costophrenic angle blunting [CUI:C0742855, counts:4542, 4542]
├── vascular redistribution [CUI:C0239041, counts:367, 499]
│   │   └── central vascular redistribution [counts:132, 132]
├── hilar enlargement [CUI:C1698506, counts:1170, 4895]
│   │   ├── adenopathy [CUI:C0149711, counts:350, 350]
│   │   └── vascular hilar enlargement [counts:3192, 3375]
│   │       └── pulmonary artery enlargement [CUI:C2072932, counts:183, 183]
├── hilar congestion [CUI:C0582411, counts:1047, 1047]
├── cardiomegaly [CUI:C0018800, counts:10527, 10527]
├── pericardial effusion [CUI:C0031039, counts:56, 56]
├── kerley lines [CUI:C0239019, counts:69, 69]
├── dextrocardia [CUI:C0011813, counts:14, 14]
├── right sided aortic arch [CUI:C0035615, counts:18, 18]
├── aortic atheromatosis [CUI:C1096249, counts:1813, 1813]
├── aortic elongation [counts:7724, 10154]
│   │   ├── descendent aortic elongation [CUI:C4476542, counts:736, 736]
│   │   ├── ascendent aortic elongation [CUI:C3889085, counts:154, 154]
│   │   ├── aortic button enlargement [CUI:C1851119, counts:368, 368]
│   │   └── supra aortic elongation [counts:1172, 1172]
├── aortic aneurysm [CUI:C0003486, counts:33, 33]
├── mediastinal enlargement [CUI:C2021206, counts:272, 5589]
```

```
│   ├── superior mediastinal enlargement [CUI:C4273001, counts:610, 2488]
│   │   ├── goiter [CUI:C0018021, counts:706, 706]
│   │   └── supra aortic elongation [counts:1172, 1172]
│   ├── descendent aortic elongation [CUI:C4476542, counts:736, 736]
│   ├── ascendent aortic elongation [CUI:C3889085, counts:154, 154]
│   ├── aortic aneurysm [CUI:C0003486, counts:33, 33]
│   ├── mediastinal mass [CUI:C0240318, counts:234, 234]
│   └── hiatal hernia [CUI:C3489393, counts:1672, 1672]
├── tracheal shift [counts:600, 600]
├── mass [CUI:C2603353, counts:45, 1026]
│   ├── mediastinal mass [CUI:C0240318, counts:234, 234]
│   ├── breast mass [CUI:C0024103, counts:1, 1]
│   ├── pleural mass [CUI:C1709576, counts:2, 2]
│   ├── pulmonary mass [CUI:C0149726, counts:654, 654]
│   └── soft tissue mass [CUI:C0457196, counts:90, 90]
├── esophagic dilatation [CUI:C0192389, counts:3, 3]
├── azygoesophageal recess shift [counts:47, 47]
├── pericardial effusion [CUI:C0031039, counts:56, 56]
├── mediastinic lipomatosis [CUI:C1333298, counts:289, 289]
├── thoracic cage deformation [CUI:C4538889, counts:166, 8702]
│   ├── scoliosis [CUI:C0036439, counts:5610, 5610]
│   ├── kyphosis [CUI:C2115817, counts:2631, 2631]
│   ├── pectum excavatum [CUI:C2051831, counts:185, 185]
│   ├── pectum carinatum [CUI:C2939416, counts:30, 30]
│   └── cervical rib [CUI:C0158779, counts:80, 80]
├── vertebral degenerative changes [CUI:C4290224, counts:2974, 4298]
│   └── vertebral compression [CUI:C0262431, counts:236, 1324]
│       └── vertebral anterior compression [counts:1088, 1088]
├── lytic bone lesion [CUI:C0476382, counts:56, 56]
├── sclerotic bone lesion [CUI:C4315325, counts:397, 450]
│   └── blastic bone lesion [CUI:C2203581, counts:53, 53]
├── costochondral junction hypertrophy [counts:98, 98]
├── sternoclavicular junction hypertrophy [counts:5, 5]
├── axial hyperostosis [CUI:C1400000, counts:184, 184]
├── osteopenia [CUI:C0029453, counts:374, 374]
├── osteoporosis [CUI:C0029456, counts:164, 164]
├── non axial articular degenerative changes [counts:236, 236]
├── subacromial space narrowing [counts:34, 34]
├── fracture [CUI:C0016658, counts:14, 3367]
│   ├── clavicle fracture [CUI:C0159658, counts:271, 271]
│   ├── humeral fracture [CUI:C0020162, counts:285, 285]
│   ├── vertebral fracture [CUI:C0080179, counts:190, 190]
│   └── rib fracture [CUI:C0035522, counts:533, 2607]
│       └── callus rib fracture [CUI:C0006767, counts:2074, 2074]
├── gynecomastia [CUI:C0018418, counts:445, 445]
├── hiatal hernia [CUI:C3489393, counts:1672, 1672]
├── Chilaiditi sign [CUI:C3178780, counts:14, 14]
├── hemidiaphragm elevation [CUI:C2073707, counts:1702, 1702]
├── diaphragmatic eventration [CUI:C0011981, counts:770, 770]
├── tracheostomy tube [CUI:C0184159, counts:1839, 1839]
├── endotracheal tube [CUI:C0336630, counts:2456, 2456]
├── NSG tube [counts:5827, 5827]
├── chest drain tube [CUI:C0008034, counts:607, 607]
├── ventriculoperitoneal drain tube [CUI:C0162702, counts:33, 33]
├── gastrostomy tube [CUI:C0150595, counts:9, 9]
├── nephrostomy tube [CUI:C0184149, counts:3, 3]
├── double J stent [CUI:C0441293, counts:11, 11]
├── catheter [CUI:C0085590, counts:63, 6167]
│   └── central venous catheter [CUI:C1145640, counts:444, 6104]
```

```
│       │     ├── central venous catheter via subclavian vein [CUI:C0398281, counts:1712, 1712]
│       │     ├── central venous catheter via jugular vein [CUI:C0398278, counts:3299, 3299]
│       │     ├── reservoir central venous catheter [CUI:C2026143, counts:466, 466]
│       │     └── central venous catheter via umbilical vein [CUI:C0398284, counts:183, 183]
│    ├── electrical device [counts:26, 4286]
│    │    ├── dual chamber device [CUI:C2732817, counts:955, 955]
│    │    ├── single chamber device [CUI:C2733312, counts:529, 529]
│    │    ├── pacemaker [CUI:C0030163, counts:2614, 2614]
│    │    └── dai [CUI:C0972395, counts:162, 162]
│    ├── artificial heart valve [CUI:C1399223, counts:488, 992]
│    │    ├── artificial mitral heart valve [CUI:C0869752, counts:299, 299]
│    │    └── artificial aortic heart valve [CUI:C0869748, counts:205, 205]
│    ├── surgery [counts:161, 7371]
│    │    ├── metal [CUI:C0025552, counts:1113, 5225]
│    │    │    ├── osteosynthesis material [CUI:C0016642, counts:530, 530]
│    │    │    ├── sternotomy [CUI:C0185792, counts:2056, 2056]
│    │    │    └── suture material [CUI:C4305366, counts:1526, 1526]
│    │    ├── bone cement [CUI:C0005934, counts:8, 8]
│    │    ├── prosthesis [CUI:C0175649, counts:17, 566]
│    │    │    ├── humeral prosthesis [CUI:C0745058, counts:117, 117]
│    │    │    ├── mammary prosthesis [CUI:C0917968, counts:382, 382]
│    │    │    └── endoprosthesis [CUI:C0005846, counts:19, 50]
│    │    │         └── aortic endoprosthesis [counts:31, 31]
│    │    ├── surgery breast [CUI:C3714726, counts:278, 729]
│    │    │    └── mastectomy [CUI:C0024881, counts:451, 451]
│    │    ├── surgery neck [CUI:C0185773, counts:442, 442]
│    │    ├── surgery lung [CUI:C0038903, counts:213, 213]
│    │    ├── surgery heart [CUI:C0018821, counts:14, 14]
│    │    └── surgery humeral [CUI:C0186326, counts:13, 13]
├── abnormal foreign body [CUI:C0016542, counts:49, 49]
└── external foreign body [counts:34, 34]
```

A.1.2 Differential Diagnoses Hierarchy

SJ-PadChest contains a total of 27,726 differential diagnoses (5,472 were manually annotated from 27,593 study reports).

label [CUI, study counts, study counts with child labels]
differential diagnosis

```
├── pneumonia [CUI:C0032285, counts:5778, 5934]
│    └── atypical pneumonia [CUI:C1412002, counts:156, 156]
├── tuberculosis [CUI:C0041296, counts:176, 824]
│    └── tuberculosis sequelae [CUI:C0494132, counts:648, 648]
├── lung metastasis [CUI:C0153676, counts:251, 251]
├── lymphangitis carcinomatosa [CUI:C0238258, counts:30, 30]
├── lepidic adenocarcinoma [CUI:C4049711, counts:10, 10]
├── pulmonary fibrosis [CUI:C0034069, counts:751, 991]
│    ├── post radiotherapy changes [CUI:C1320687, counts:196, 196]
│    └── asbestosis signs [CUI:C0003949, counts:44, 44]
├── emphysema [CUI:C0034067, counts:958, 958]
├── COPD signs [CUI:C0024117, counts:14557, 14557]
├── heart insufficiency [CUI:C0018801, counts:2339, 2339]
├── respiratory distress [CUI:C0476273, counts:53, 53]
├── pulmonary hypertension [CUI:C0020542, counts:95, 129]
│    ├── pulmonary artery hypertension [CUI:C2973725, counts:28, 28]
│    └── pulmonary venous hypertension [CUI:C4477098, counts:6, 6]
├── pulmonary edema [CUI:C0034063, counts:1515, 1515]
```

└── bone metastasis [CUI:C0153690, counts:135, 135]

A.1.3 Anatomical Locations

A total of 249,469 differential diagnoses were annotated (55,472 manually tagged in 27,593 study reports).

label [CUI, study counts, study counts with child labels]
localization
├── extracorporal [CUI:C0424529, counts:0, 0]
├── cervical [CUI:C0920882, counts:992, 992]
├── soft tissue [CUI:C0225317, counts:413, 3567]
│   ├── subcutaneous [CUI:C0443315, counts:330, 330]
│   ├── axilar [CUI:C0004454, counts:863, 863]
│   └── pectoral [CUI:C0230111, counts:1081, 1961]
│   └── nipple [CUI:C0028109, counts:880, 880]
├── bone [CUI:C0262950, counts:1472, 12862]
│   ├── shoulder [CUI:C0037004, counts:287, 2149]
│   │   ├── acromioclavicular joint [CUI:C0001208, counts:69, 69]
│   │   ├── rotator cuff [CUI:C0085515, counts:67, 67]
│   │   ├── supraspisnous [CUI:C0225001, counts:43, 43]
│   │   └── humerus [CUI:C0020164, counts:1051, 1683]
│   │   ├── humeral head [CUI:C0223683, counts:430, 430]
│   │   ├── humeral neck [CUI:C0448034, counts:88, 88]
│   │   └── glenohumeral joint [CUI:C0225063, counts:114, 114]
│   ├── clavicle [CUI:C0008913, counts:388, 388]
│   ├── scapula [CUI:C0036277, counts:171, 171]
│   ├── costoesternal [CUI:C0450216, counts:36, 36]
│   ├── column [CUI:C0037949, counts:1431, 3783]
│   │   ├── intersomatic space [CUI:C0223088, counts:52, 52]
│   │   ├── dorsal vertebrae [CUI:C0039987, counts:1979, 1979]
│   │   ├── cervical vertebrae [CUI:C3665420, counts:100, 100]
│   │   └── paravertebral [CUI:C0442150, counts:221, 221]
│   └── rib [CUI:C0035561, counts:3756, 4863]
│   ├── anterior rib [CUI:C4323264, counts:320, 320]
│   ├── posterior rib [CUI:C4323265, counts:750, 750]
│   └── rib cartilage [CUI:C0222787, counts:37, 37]
├── hemithorax [CUI:C0934569, counts:5119, 5119]
├── extrapleural [CUI:C0442091, counts:73, 73]
├── extrapulmonary [counts:46, 46]
├── pleural [CUI:C0032225, counts:13935, 13935]
├── subpleural [CUI:C0225775, counts:115, 115]
├── fissure [CUI:C0458078, counts:734, 1318]
│   ├── minor fissure [CUI:C0734040, counts:486, 486]
│   └── major fissure [CUI:C4253583, counts:98, 98]
├── lobar [CUI:C0225752, counts:2413, 14864]
│   ├── upper lobe [CUI:C0225756, counts:1222, 6720]
│   │   ├── left upper lobe [CUI:C1261076, counts:1574, 2477]
│   │   │   └── lingula [CUI:C0225740, counts:903, 903]
│   │   └── right upper lobe [CUI:C1261074, counts:3021, 3021]
│   ├── lower lobe [CUI:C0225758, counts:792, 4144]
│   │   ├── left lower lobe [CUI:C1261077, counts:1994, 1994]
│   │   └── right lower lobe [CUI:C1261075, counts:1358, 1358]
│   └── middle lobe [CUI:C4281590, counts:1587, 1587]
├── subsegmental [CUI:C0929165, counts:1850, 1850]
├── bronchi [CUI:C0006255, counts:3315, 3315]
├── peribronchi [CUI:C0225607, counts:114, 114]
├── lung field [CUI:C0225759, counts:819, 78888]
│   ├── pleural [CUI:C0032225, counts:13935, 13935]

```
│   ├── subpleural [CUI:C0225775, counts:115, 115]
│   ├── major fissure [CUI:C4253583, counts:98, 98]
│   ├── subsegmental [CUI:C0929165, counts:1850, 1850]
│   ├── bronchi [CUI:C0006255, counts:3315, 3315]
│   ├── peribronchi [CUI:C0225607, counts:114, 114]
│   ├── upper lung field [CUI:C0929227, counts:494, 10442]
│   │   └── upper lobe [CUI:C0225756, counts:1222, 9948]
│   │       ├── left upper lobe [CUI:C1261076, counts:1574, 1574]
│   │       ├── right upper lobe [CUI:C1261074, counts:3021, 3021]
│   │       ├── apical [CUI:C0734296, counts:3884, 3884]
│   │       └── suprahilar [CUI:C0934260, counts:247, 247]
│   ├── middle lung field [CUI:C0929434, counts:1591, 13539]
│   │   ├── aortopulmonary window [CUI:C1282038, counts:113, 113]
│   │   ├── hilar [CUI:C0205150, counts:6678, 11349]
│   │   │   ├── pulmonary artery [CUI:C0034052, counts:215, 215]
│   │   │   ├── hilar bilateral [counts:690, 690]
│   │   │   └── perihilar [CUI:C0225702, counts:3766, 3766]
│   │   └── minor fissure [CUI:C0734040, counts:486, 486]
│   └── lower lung field [counts:1352, 34661]
│       ├── basal [CUI:C1282378, counts:9632, 9632]
│       ├── lower lobe [CUI:C0225758, counts:792, 4144]
│       │   ├── left lower lobe [CUI:C1261077, counts:1994, 1994]
│       │   └── right lower lobe [CUI:C1261075, counts:1358, 1358]
│       ├── middle lobe [CUI:C4281590, counts:1587, 1587]
│       ├── infrahilar [counts:675, 675]
│       ├── lingula [CUI:C0225740, counts:903, 903]
│       ├── supradiaphragm [counts:19, 19]
│       ├── diaphragm [CUI:C0011980, counts:3798, 3798]
│       ├── infradiaphragm [counts:2934, 2934]
│       ├── cardiophrenic angle [counts:287, 287]
│       └── costophrenic angle [CUI:C0230151, counts:5805, 9330]
│           ├── right costophrenic angle [CUI:C0504099, counts:1445, 1445]
│           ├── left costophrenic angle [CUI:C0504100, counts:2019, 2019]
│           └── bilateral costophrenic angle [counts:61, 61]
├── central [CUI:C0205099, counts:5488, 5488]
├── mediastinum [CUI:C0025066, counts:2686, 39134]
│   ├── superior mediastinum [CUI:C0230147, counts:645, 8449]
│   │   ├── carotid artery [CUI:C0007272, counts:4, 4]
│   │   ├── brachiocephalic veins [CUI:C0006095, counts:203, 203]
│   │   ├── supra aortic [counts:1123, 1123]
│   │   ├── aortic button [CUI:C0003489, counts:348, 348]
│   │   ├── superior cave vein [CUI:C3165182, counts:3663, 3663]
│   │   └── subclavian vein [CUI:C0038532, counts:2463, 2463]
│   ├── lower mediastinum [CUI:C1261193, counts:200, 18497]
│   │   ├── anterior mediastinum [CUI:C0230148, counts:64, 162]
│   │   │   └── thymus [CUI:C0040113, counts:98, 98]
│   │   ├── middle mediastinum [CUI:C0230149, counts:32, 15740]
│   │   │   └── cardiac [CUI:C1522601, counts:15663, 15708]
│   │   │       └── coronary [CUI:C1522318, counts:45, 45]
│   │   └── posterior mediastinum [CUI:C0230150, counts:16, 2395]
│   │       └── retrocardiac [counts:2379, 2379]
│   └── aortic [CUI:C0003483, counts:9502, 9502]
├── paratracheal [CUI:C0442143, counts:480, 480]
├── airways [CUI:C0458827, counts:18, 8685]
│   ├── tracheal [CUI:C0040578, counts:5352, 5352]
│   └── bronchi [CUI:C0006255, counts:3315, 3315]
├── esophageal [CUI:C1522619, counts:239, 239]
├── paramediastinum [counts:70, 70]
├── paracardiac [counts:188, 188]
```

├── epigastric [counts:42, 42]
├── gastric chamber [CUI:C3714551, counts:887, 887]
├── hypochondrium [CUI:C0230186, counts:193, 455]
│   ├── right hypochondrium [CUI:C0738590, counts:122, 196]
│   │   └── gallbladder [CUI:C0016976, counts:74, 74]
│   └── left hypochondrium [CUI:C0738591, counts:66, 66]
├── right [CUI:C0444532, counts:23628, 23628]
├── left [CUI:C0443246, counts:17133, 17133]
└── bilateral [CUI:C0238767, counts:11034, 15982]
    ├── diffuse bilateral [counts:1636, 1636]
    └── basal bilateral [counts:3312, 3312]

B   Regular expressions used to extract spatial concepts from Spanish reports

Regular expression	Spatial Concept
\bsen\scost.*\sderech	right costophrenic angle
\bsen\scost.*\sderech	right costophrenic angle
\bsen\scost.*\sizq	left costophrenic angle
\bsen\scost.*\sbil	bilateral costophrenic angle
\bsen\scost	costophrenic angle
\bsupradiafrag	supradiaphragm
\bhemidiafrag\bdiafrag	diaphragm
\binfradiafrag	infradiaphragm
\blobul\b	lobar
\bsubsegment	subsegmental
\bperibronqu	peribronchi
\bbronqui\bendobronqui\bbronc	bronchi
\blingul	lingula
(\bls\s\lobul\ssup)	upper lobe
(\bli\s\lobul\sinf)	lower lobe
(\blsi\bls\sizq\lobul\ssup.*sizq)	left upper lobe
(\blsd\bls\sderech\lobul\ssup.*sderec)	right upper lobe
(\blmd\blm\b)	middle lobe
\bretrocardi	retrocardiac
\blii	left lower lobe
\blid	right lower lobe
\bcamp.*\sinfer	lower lung field
\bcamp.*\ssup	upper lung field
\bcamp.*\smed	middle lung field
\bcamp.*\spulm	lung field
\baxil	axilar
(\bcervical\bcuell\b)	cervical
(\bvertic\bapic\bapex)	apical
\bbas	basal
\bcentral	central
\bcostal	rib
\bcostal anterior	anterior rib
\bcostal posterior	posterior rib
\bextrapleur	extrapleural
\bpleur	pleural
\bsubpleur	subpleural
(\bfundus gastric\bcam gastric)	gastric chamber
\bhili.*\sbila	hilar bilateral
\bhili	hilar
(\bperihili\bparahili)	perihilar

Continued on next column

32

Continued from previous column

Regular expression	Spatial Concept
\bsuprahili	suprahilar
\binfrahili	infrahilar
\bmediastin sup	superior mediastinum
\bmediastin anterior	anterior mediastinum
\bmediastin medi	middle mediastinum
\bmediastin posterior	posterior mediastinum
\bmediastin infer	lower mediastinum
\bmediastin	mediastinum
\bparamediastin	paramediastinum
\btim	thymus
\bventan aort	aortopulmonary window
(\bcolumn.*\sdorsal|\bvertebr.*\sdorsal)	dorsal vertebrae
(\bcolumn.*\scerv|\bvertebr.*\scerv)	cervical vertebrae
(\bcolumn.*\sax|\bcolumn.*\svert|\bvertebr)	column
\bparavertebr	paravertebral
\bhipocondri	hypochondrium
\bhipocondri derech	right hypochondrium
\bhipocondri izq	left hypochondrium
\bhues|\bose	bone
\bhumer	humerus
\bcabez.*\shumer	humeral head
\bcuell.*\shumer	humeral neck
\bglen	glenohumeral joint
\bacromi.*clavi	acromioclavicular joint
\bclavicul	clavicle
\bescap	scapula
\bsupraespin	supraspisnous
\bhombr	shoulder
\bcostoesternal	costoesternal
\bcartilag cost	rib cartilage
\bcardiofren	cardiophrenic angle
\bsiluet cardi|\bcardi|\bcorazo	cardiac
\bparacardi	paracardiac
\bcoronar	coronary
\bsupraaort	supra aortic
\bboton aort	aortic button
\baort	aortic
\barteri pulmon|p̂ulmonar\b	pulmonary artery
\bcav\ssup	superior cave vein
\bsubclavi	subclavian vein
\bbraquio	brachiocephalic veins
\bcarotid	carotid artery
\bsubcutan	subcutaneous
\bextracorp	extracorporal
\bmanguit.*\srota	rotator cuff
\bcisur.*\smenor\b	minor fissure
\bcisur.*\smayor\b	major fissure
\bcisur\b	fissure
\bintersomat	intersomatic space
\bextrapulmo	extrapulmonary
\bc\d\b	cervical vertebrae
\bd\d?\d\b	dorsal vertebrae
\bl\d\b	lumbar vertebrae
\bepigastri	epigastric
\bparatraque	paratracheal
\bvia.*\saere	airways

Continued on next column

| Continued from previous column	
Regular expression	**Spatial Concept**
\btraque|\bendotraque|\bintratraque	tracheal
\besofag	esophageal
\bbland	soft tissue
\bpectoral|\bmama|\bginecomas	pectoral
\bpezon|\bmamil	nipple
(\bcolecist|colelit|vesicul.*\bbil)	gallbladder
\bhemit	hemithorax
\bderech\b	right
\bizq	left
\bdifus.*\sbilat	diffuse bilateral
\bbilat.*\sdifus	diffuse bilateral
(\bbilat|\bamb.*hemit|\bamb.*camp)	bilateral
\bbibas	basal bilateral
	Concluded

C Dataset example

Fig. 18 shows an example of an study with two x-ray images. The associated fields in Tab. 10 are shared for both projections included in the same study report.
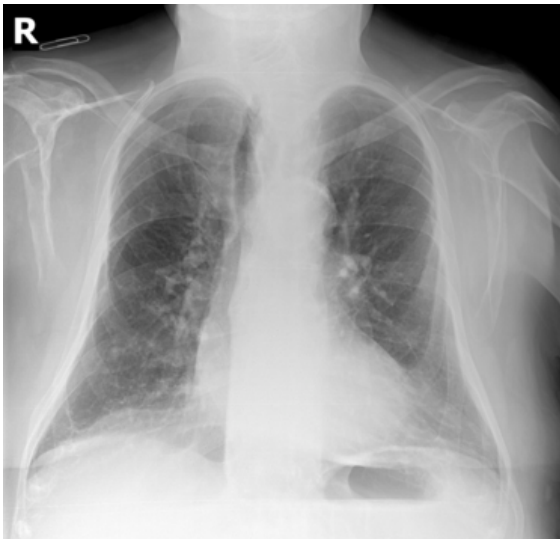
(a) P-A
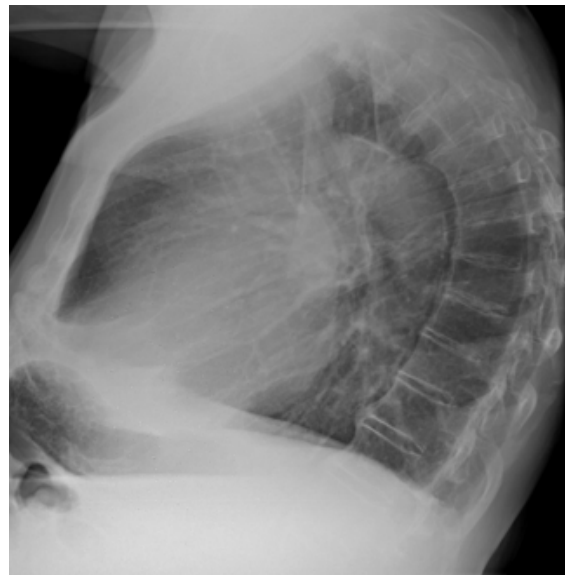
(b) L

Figure 18: Example x-ray images of the study with two projections which fields are given in Tab. 10.

ImageID	135803415504923515076821959678074435083_fzis7b.png
ImageDir	2
StudyDate_DICOM	2015-09-14
StudyID	20150914, 135803415504923515076821959678074435083
PatientID	313572750430997347502932654319389875966
PatientBirth	1929
ReportID	4991845
Report	cambi pulmonar cronic sever . sign fibrosis bibasal . sutil infiltr pseudonodul milimetr vidri deslustr localiz bas . cifosis sever .
MethodLabel	Physician
Labels	['pulmonary fibrosis', 'chronic changes', 'kyphosis', 'pseudonodule', 'ground glass pattern']
LabelsLocalizationsBySentence	[['pulmonary fibrosis', 'loc basal bilateral'], ['chronic changes'], ['kyphosis'], ['pseudonodule', 'ground glass pattern', 'loc basal']]
LabelCUIS	['C0034069' 'C0742362' 'C2115817' 'C3544344']
LocalizationsCUIS	['C1282378']
PatientSex_DICOM	M
ViewPosition_DICOM	POSTEROANTERIOR
Projection	PA
MethodProjection	manual
Modality_DICOM	
Manufacturer_DICOM	ImagingDynamicsCompanyLtd
PhotometricInterpretation_DICOM	MONOCHROME2
PixelRepresentation_DICOM	
PixelAspectRatio_DICOM	
SpatialResolution_DICOM	
BitsStored_DICOM	12
WindowCenter_DICOM	2155
WindowWidth_DICOM	2880
Rows_DICOM	3572
Columns_DICOM	3732
XRayTubeCurrent_DICOM	320
Exposure_DICOM	3
ExposureInuAs_DICOM	3200
ExposureTime_DICOM	10
RelativeXRayExposure_DICOM	

Table 10: Fields corresponding to the study which images are shown in Fig. 18.