Chest ImaGenome Dataset for Clinical Reasoning

Joy T. Wu[1], Nkechinyere N. Agu[2], Ismini Lourentzou[3], Arjun Sharma[4], Joseph A. Paguio[5], Jasper S. Yao[5], Edward C. Dee[6], William Mitchell[4], Satyananda Kashyap[1], Andrea Giovannini[1], Leo A. Celi[4], Mehdi Moradi[1]

[1]IBM Almaden Research Center, San Jose, CA 95120, USA
[2]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[3]Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[4]MIT Critical Data, Cambridge, MA 02139, USA
[5]Albert Einstein Healthcare Network-Philadelphia Campus, PA 19141, USA
[6]Harvard Medical School, Boston, MA 02115, USA

Abstract

Despite the progress in automatic detection of radiologic findings from Chest X-Ray (CXR) images in recent years, a quantitative evaluation of the explainability of these models is hampered by the lack of locally labeled datasets for different findings. With the exception of a few expert-labeled small-scale datasets for specific findings, such as pneumonia and pneumothorax, most of the CXR deep learning models to date are trained on global "weak" labels extracted from text reports, or trained via a joint image and unstructured text learning strategy. Inspired by the Visual Genome effort in the computer vision community, we constructed the first Chest ImaGenome dataset with a scene graph data structure to describe $242,072$ images. Local annotations are automatically produced using a joint rule-based natural language processing (NLP) and atlas-based bounding box detection pipeline. Through a radiologist constructed CXR ontology, the annotations for each CXR are connected as an anatomy-centered scene graph, useful for image-level reasoning and multimodal fusion applications. Overall, we provide: i) $1,256$ combinations of relation annotations between 29 CXR anatomical locations (objects with bounding box coordinates) and their attributes, structured as a scene graph per image, ii) over $670,000$ localized comparison relations (for improved, worsened, or no change) between the anatomical locations across sequential exams, as well as ii) a manually annotated gold standard scene graph dataset from 500 unique patients.

Introduction

Chest X-Rays (CXR) are among the commonly ordered radiology exams, mostly for screening but also for diagnostic purposes. Recently, multiple large CXR imaging datasets have been released by the research community [12, 20, 22, 42]. These can be used to develop automatic abnormality detection or report generation algorithms. For detecting specific abnormalities from images, natural language processing (NLP) algorithms have been used to extract "weak" global image-level labels (CXR abnormalities) from the associated CXR reports [5, 20, 39, 45]. For automatic report generation, self-supervised joint text and image architectures [27, 31, 43, 49, 50], first inspired by the image captioning related work in the non-medical domain [14, 24, 34, 41, 47], have been used to produce preliminary free-text radiology reports. However, both approaches lack rigorous localization assessment for explainability, namely whether the model attended to the relevant anatomical location(s) for predictions. This missing feature is critical for clinical applications. The joint image and text learning strategy is also known to learn heavy language priors from the text reports without having learned to interpret the imaging features [1, 36]. Furthermore, even though there exist architectures suitable for comparing imaging changes [28, 29], limited work has focused on automatically deriving

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

comparison relations between exams from large datasets, for the purpose of training imaging models that can track progress for a wide variety of CXR findings or diseases.

To the best of our knowledge, no prior work in CXR has attempted to automatically extract relations between CXR attributes (labels) from reports and their anatomical locations (objects with bounding box coordinates) on the images as documented by the reporting radiologists, nor has there been any localized relation annotations between sequential CXR exams. Research on these two topics is valuable because radiology reports in effect are records of radiologists' complex clinical reasoning processes, where the anatomical location of observed imaging abnormalities is often used to narrow down on potential diagnoses, as well as for integrating information from other clinical modalities (e.g., CT findings, labs, etc) at the anatomical levels. Sequential exams are also routinely used by bedside clinicians to track patients' clinical progress after being started on different management paths. Therefore, documentations comparing sequential exams are prevalent in CXR reports and are clinically meaningful relations to learn about. Automatically extracting radiology knowledge graphs and disease progression information from reports will help improve explainability evaluation and widen downstream clinical applications for CXR imaging algorithm development.

Many algorithms for object detection and domain-knowledge-driven reasoning require a starting dataset that has localized labels on the images and meaningful relationships between them. In the non-medical domain, large locally labeled graph datasets (e.g., Visual Genome dataset [26]) have enabled the development of algorithms that can integrate both visual and textual information and derive relationships between observed objects in images [30, 46, 48]. In addition, they have spurred a whole domain of research in visual question answering (VQA) and visual dialogue (VD), with the aim of developing interactive AI algorithms capable of reasoning over information from multiple sources [3, 9, 11]. These location, relation and semantics aware systems aim to capture important elements in image data in relation to complex human languages, in order to conversationally interact with humans about the visual content. In the medical domain, such systems may help with automatic image and text information retrieval tasks from databases, or improve end-user trust by allowing clinicians to interactively question trained models to assess the consistency of predictions.

In this paper, we present the Chest ImaGenome dataset, a large multi-modal (text and images) chronologically ordered scene graph dataset for frontal Chest X-ray (CXR) images. This dataset is an important step towards addressing the missing link of large locally labeled graph datasets in the medical imaging domain. The goal for releasing this dataset is to spur the development of algorithms that more closely reflect radiology experts' reasoning processes. In addition, automatically describing localized imaging features in recognized medical semantics is the first step towards connecting potentially predictive pixel-level features from medical images with the rest of the digital patient records and external medical ontologies. These connections could aid both the development of anatomically relevant multi-modal fusion models and the discovery of localized imaging fingerprints, i.e., patterns predictive of patient outcomes. Through **PhysioNet's credentialed access** (see **license**), we make the first Visual Genome-like graph dataset in the CXR domain accessible for the research community.

Related work: A few CXR datasets have localized abnormality annotations [13, 21, 38] that are curated manually. These are high quality gold standard ground truth datasets but tend to be smaller in scale (< 30,000 images) and have a narrow coverage, with typically only 1-2 labels. In addition, since most labeling efforts only have abnormality semantics attached, no direct relationships with the affected anatomical locations are available.

Two recent CXR datasets have labels for anatomies described in the reports. In [10], a small manually annotated dataset (2000 reports) included 10 abnormalities that are individually associated with 29 unique spatial locations (anatomies) at the report level. Another CXR dataset has automatically extracted abnormality and anatomy labels as disconnected concepts that are only correlated at the study level from 160,000 reports using a supervised NLP algorithm [5]. This was trained on a smaller set of manually annotated data. Neither datasets contain localized annotations for the associated CXR images, nor any comparison relation annotations between sequential exams, both of which are available in the Chest ImaGenome dataset. In Table 1, we present a comparison of our Chest ImagGenome dataset with other datasets available in the literature.

Table 1: Summary of existing Chest X-ray datasets

Dataset	Annotation Level	Annotation Method	Num Labels	Anatomy Labeled	Graph	Dataset Size	Temporal Labels	Reports
SIIM-ACR Pneumothorax Segmentation [13]	Segmentation	Manual + augmented	1	No	No	12,047	No	No
RSNA Pneumonia Detection Challenge [38]	Bounding Boxes	Manual	1	No	No	30,000	No	No
Indiana University Chest X-ray collection [12]	Global	Automated	10	No	No	3,813	No	Yes
NIH CXR dataset [42]	Global	Automated	14	No	No	112,120	No	No
PLCO [40]	Global	Automated	24	Yes	No	236,000	Yes	No
Stanford CheXpert [20]	Global	Automated	14	No	No	224,316	No	No
MIMIC-CXR [22]	Global	Automated	14	No	No	377,110	No	Yes
Dutta [10]	Global	Manual	10	Yes	Yes	2,000	No	Yes
PadChest [5]	Global	Manual + automated	297	Yes	No	160,868	No	Yes
Montgomery County Chest X-ray [21]	Segmentation	Manual	1	Yes	No	138	No	No
Shenzen Hospital Chest X-ray [21]	Segmentation	Manual	1	Yes	No	662	No	No
Chest ImaGenome	Bounding Boxes	Automated	131	Yes	Yes	242,072	Yes	Yes

Methods

The Chest ImaGenome dataset was derived from the MIMIC-CXR dataset [22], which has been de-identified. This derived dataset retains the added annotations and the source image tags but not the CXR images, which users are expected to separately download from the **MIMIC-CXR database**. The institutional review boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (BIDMC)(2001-P-001699/14) both approved the use of the MIMIC database for research. All authors working with the data have individually completed required HIPAA training and been granted data access approval from PhysioNet.

Silver Dataset Construction

The Chest ImaGenome dataset construction (Supplementary Figure 3) is inspired by the Visual Genome dataset [26]. Whereas Visual Genome utilized web-based and crowd-sourced methods to manually collect annotations, the Chest ImaGenome harnessed NLP, a CXR ontology, and image segmentation techniques to automatically structure and add value to existing CXR images and their free-text reports, which were collected from radiologists in their routine workflow. We used atlas-based bounding box extraction techniques to structure the anatomies on $242,072$ frontal CXR images, anteroposterior (AP) or posteroanterior (PA) view, and used a rule-based text-analysis pipeline to relate the anatomies to various CXR attributes (finding, diseases, technical assessment, devices, etc) extracted from $217,013$ reports. Altogether, we automatically annotated $242,072$ scene graphs that locally and graphically describe the frontal images associated with these reports (one report can have one or more frontal images). Our goal is not only to locally label attributes relevant for key anatomical locations on the CXR images, but also to extract documented radiology knowledge from a large corpus of CXR reports to aid future semantics-driven and multi-modal clinical reasoning works.

Table 2 describes the parallels between the Chest ImaGenome and Visual Genome datasets. The key differences are in the construction methodology, the currently much smaller range of possible objects and attributes (due to having only the CXR imaging modality), and the introduction of comparison relations between sequential images in the Chest ImaGenome dataset. We define the nodes and edges in the graph (Supplementary Table 6) based on clinical relevance and resources in the context for medical imaging exams like CXRs. In addition, two **key assumptions** are made in the construction of the Chest ImaGenome dataset:

1) CXR imaging observations can be normalized to relationships between the visualized anatomical locations (object nodes) and the abnormalities, devices or other CXR descriptions (attribute nodes) that the locations contain. Thus, the variety of detected objects is confined by the granularity of anatomical location detection on images and from reports.

2) The exam timestamps in the original MIMIC-CXR dataset can be used to chronologically order the CXR exams from the same patient within the original MIMIC CXR dataset's collection period and there are minimal missing exams for each patient. This is based on discussions with the MIMIC team and MIMIC-CXR's documented data collection strategy. The original data curators included all CXR exams in the radiology imaging archives for patients who were at any time point admitted to the BIDMC's Emergency Department within a continuous 2-year-period. Therefore, we related any comparison descriptions (normalized to 'improved', 'worsened' and 'no change') of attribute(s) in different anatomical location(s) to the same anatomical location(s) on the exam image(s) immediately before the current exam. Clinically, the extracted comparison relations are intended to allow longitudinal modeling of disease progression for different CXR anatomies.

The construction of the Chest ImaGenome dataset builds on the works of [44, 45]. In summary, the text pipeline [45] first sections the report and retains only the finding and impression sentences, and

Table 2: Parallels between the Chest ImaGenome and Visual Genome datasets.

Element	Chest ImaGenome	Visual Genome
Scene	One frontal CXR image in the current dataset.	One (non-medical) everyday life image.
Questions	For now, there is only one question per CXR, which is taken from the patient history (i.e., reason for exam) section from each CXR report.	One or more questions that the crowd source annotators decided to ask about the image where the information from each question and the image should allow another annotator to answer it.
Answers	N/A currently. However, report sentences are biased towards answering the question asked in the reason for exam sentence;hence, the knowledge graph we extract from each report should contain the answer(s).	This was collected as answer(s) to the corresponding question(s) asked of the image.
Sentences (Region descriptions)	Sentences from the finding and impression sections of a CXR report describing the exam as collected from radiologists in their routine radiology workflow.	True natural language descriptive sentences about the image collected from crowd-sourced everyday annotators.
Objects (nodes)	Anatomical structures or locations that have bounding box coordinates on the associated CXR image, and is indexed to the UMLS ontology [4].	The people and physical objects with bounding box coordinates on the image and indexed to WordNet ontology [32].
Attributes (nodes)	Descriptions that are true for different anatomical structures visualized on the CXR image (e.g., There is a right upper lung [object] opacity [attribute]), indexed to the UMLS ontology [4]. No Bbox coordinates.	Various descriptive properties of the objects in the image (e.g., The shirt [object] is blue [attribute]), indexed to WordNet ontology [32]. No Bbox coordinates.
Relations: object and attribute	The relationship(s) between an anatomical object and its attribute(s) from the same CXR image (e.g., There is a [relation] right upper lung [object] opacity [attribute]).	The relationship(s) between an object and its attribute(s) from the same image (e.g., The shirt [object] is [relation] blue [attribute]).
Relations: object and object	The comparison relationship (index to UMLS [4]) between the same anatomical object from two sequential CXR images for the same patient (e.g., There is a new [relation] right lower lobe [current and previous anatomical objects] atelectasis [attribute]).	The relationship (indexed to WordNet [32]) between objects in the same image (e.g., The boy [object 1] is beside [relation] the bus [object 2]).
Relations: parent and child	To make the graph for each image logically consistent and correct as learnable and consumable radiology knowledge, affirmed parent-child relations between nodes are embedded in the scene graphs – i.e., if a child attribute is related to an object, then its parent would be too (e.g., if right lung has consolidation [child], then it also has lung opacity [parent]).	N/A due to different graph construction strategy and goals. The annotators were asked to describe any (but not all) relations they observe in an image.
Scene graph	Constructed from the objects, the attributes and the relationships between them for the image.	Same but the nodes and edges overall would be more varied than Chest ImaGenome for now.
Sequence*	A super-graph for a set of chronologically ordered series of exams for the same patient.	N/A, but would be a graph for a video in the non-medical context.

then utilizes a CXR concept dictionary (lexicons) to spot and detect the context (negated or affirmed) of 271 different CXR related named-entities from each retained sentence. The lexicons were curated in advance by two radiologists in consensus using a concept expansion and vocabulary grouping engine [8]. A set of sentence-level filtering rules are applied to disambiguate some of the target concepts (e.g., 'collapse' mention in CXR report can be about lung 'collapse' or related to spinal fracture as in vertebral body 'collapse'). Then the named-entities for CXR labels (attributes) are associated with the name-entities for anatomical location(s) described in the same sentence with a SpaCy natural language parser [18].

Using a CXR ontology constructed by radiologists, a scene graph assembly pipeline corrected obvious attribute-to-anatomy assignment errors (e.g., lung opacity wrongly assigned to mediastinum). Finally, the attributes for each of the target anatomical regions from repeated sentences are grouped to the exam level. The result is that, from each CXR report, we extract a radiology knowledge graph where CXR anatomical locations are related to different documented CXR attribute(s). The "reason for exam" sentence(s) from each report, which contain free text information about prior patient history, are separately kept in the final scene graph JSONs. Patient history information is critical for clinical reasoning but is a piece of information that is not technically part of the "scene" for each CXR.

For detecting the anatomical "objects" on the CXR images that are associated with the extracted report knowledge graph, a separate anatomy atlas-based bounding box pipeline extracts the coordinates of those anatomies from each frontal image. This pipeline is an extension of prior work that covers additional anatomical locations in this dataset [44]. In addition, MDs manually validated or corrected the bounding boxes for $1,071$ CXR images (with and without disease, and excluded gold standard

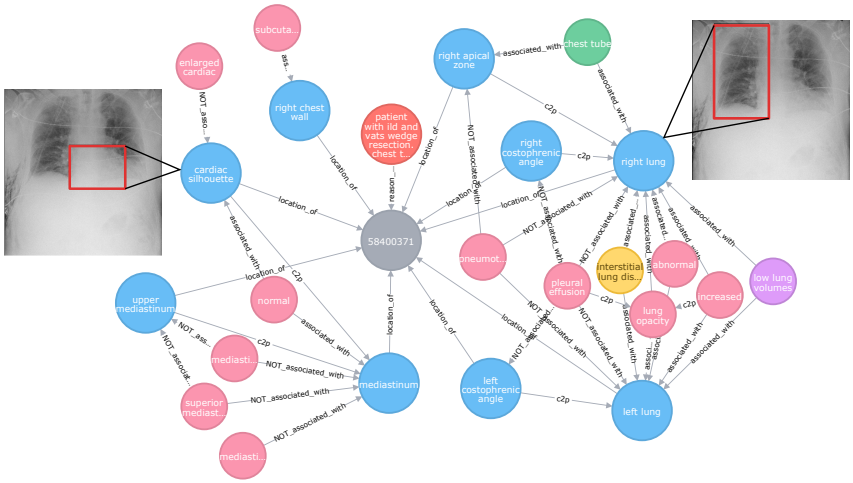

Figure 1: A radiology knowledge graph extracted for one CXR report (grey), with patient history from indication for exam (orange), anatomical locations (blue) and their associated attributes, including anatomical findings (pink), diseases (yellow), technical assessment (purple) and devices (green) nodes. The blue anatomy nodes (a.k.a. objects) also have corresponding bounding box coordinates on the CXR image, which are shown for two examples.

subjects) to train a Faster-RCNN CXR bounding box detection model, which we used to correct failed bounding boxes (too small or missing) from the initial bounding box extraction pipeline (7%). Finally, for quality assurance, we manually annotated 303 images that had missing bounding boxes for key CXR anatomies (lungs and mediastinum).

Extracting comparison relations between sequential exams at the anatomical level is another goal for the Chest ImaGenome dataset. After checking with the MIMIC team and reviewing their dataset documentation, we assume that the timestamps in the original MIMIC-CXR dataset can be used to chronologically order the exams for each patient. We then correlated all report descriptions of changes (grouped as improved, worsened, or no change) between sequential exams with the anatomical locations described at the sentence level. To extract these comparison descriptions, we used a concept expansion engine [8] to curate and group relevant comparison vocabularies used in CXR reports. These comparison relations extracted between anatomical locations from sequential CXRs are only added to the final scene graphs for every patient's second or later CXR exam(s), i.e., comparison relations described in the first study of each patient in the MIMIC-CXR dataset are not added to the Chest ImaGenome dataset.

Finally, we have mapped all object and attribute nodes and comparison relations in the dataset to a Concept Unique Identifier (CUI) in the Unified Medical Language System (UMLS) [4]. The UMLS ontology has incorporated the concepts from the Radlex ontology [31], which targets the radiology domain. Choosing UMLS to index the Chest ImaGenome dataset widens its future applications in clinical reasoning tasks, which would invariably require medical concepts and relations outside the radiology domain. An example of a CXR scene graph is shown in Figure 1.

Gold Standard Dataset Collection

In collaboration with clinicians (radiology and internal medicine M.D.'s) from multiple academic institutions, we curated a dual validated gold standard dataset to 1) evaluate the quality of the silver Chest ImaGenome dataset we automatically generated, and 2) to serve as a benchmark resource for future research using the dataset. Due to resource constraints, we created the gold standard dataset using a validation-plus-correction strategy. We randomly sampled 500 unique patients from the Chest ImaGenome dataset that had two or more sequential CXR exams. Overall, we targeted three aspects of the scene graph dataset generation process to evaluate separately: A) the object-to-attribute relations (i.e., CXR knowledge graph) extracted from individual reports, B) the object-to-object comparison relations extracted between sequential CXR reports, and C) the anatomical location detection (i.e., the bounding box extraction pipeline) for the CXR images. For details about the gold standard dataset annotation process, see Supplementary (Section 4).

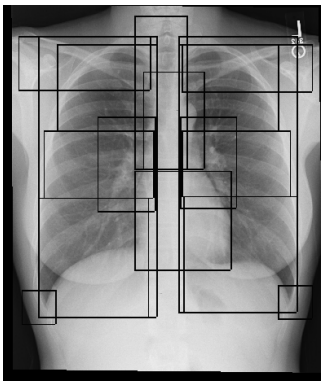

Figure 2: Sample CXR case with 17 overlaying clavicles, lung and mediastinum related anatomical bounding boxes (objects).

Data description

The Chest ImaGenome dataset is committed to the PhysioNet repository in two main directories, one for the scene graphs that are automatically generated ("silver_dataset"), and another for the 500 unique patient subset that was manually validated and corrected ("gold_dataset"). Overall, $242,072$ scene graphs were automatically derived from $217,013$ unique CXR studies. The nodes and edges in the graph are defined in detail in Supplementary Table 6. On average, 7 anatomical objects and 5 attributes are extracted from each study report. However, up to 29 anatomy objects can be detected in each CXR image with a percentage of misses $< 0.02\%$ for most objects (See Table 7 in Supplementary material). In addition, even without considering the related attribute(s), $678,543$ object-object comparison relations are extracted between anatomies across $128,468$ pairs of sequential CXR images. Detailed dataset characteristics are explained and provided in the PhysioNet repository (generate_scenegraph_statistics.ipynb). Figure 2 shows an example with 17 of the anatomical bounding boxes.

Chest ImaGenome Scene Graph JSONs

The 'silver_dataset/scene_graph.zip' file is a directory that contains multiple JSON files, one for each scene graph. Each scene graph describes one frontal Chest X-ray image. The structure for each scene graph JSON is described by components for easier explanation in Supplementary (Section 3). The first level of the JSON in Supplementary (Section 3.1) describes the patient or study level information that may not be available in the image. The fields are: 'image_id' (dicom_id in MIMIC-CXR), 'viewpoint' (AP or PA), 'patient_id' (subject_id in MIMIC-CXR), 'study_id' (study_id in MIMIC-CXR), 'gender' and 'age_decile' demographics (from MIMIC-CXR's metadata), 'reason for exam' (patient history sentence(s) from the CXR reports with age removed), 'StudyOrder' (the order of the CXR study for the patient, which is derived from chronologically ordering the DICOM timestamps), and 'StudyDateTime'; (from MIMIC's dicom metadata, which had been de-identified into the future).

For each scene graph, there are 3 separate nested fields to describe the "objects" on the CXR images, the "attributes" related to the different objects as extracted from the corresponding reports, and "relationships" to describe comparison relations between sequential CXR images for the same patient. These 3 fields are a list of dictionaries, where the format of each dictionary is modeled after the respective JSONs in the Visual Genome dataset [26].

For objects, each dictionary has the format shown in Supplementary (Section 3.2). The 'object_id' is unique across the whole dataset for the anatomical location on the particular image. Fields 'x1', 'y1', 'x2', 'y2', 'width' and 'height' are for a padded and resized 224×224 CXR frontal image, where coordinates 'x1', 'y1' are for the top left corner of the bounding box and 'x2', 'y2' are for the bottom right corner. The bounding box coordinates in the original image are denoted with 'original_*'. The remaining fields: 'bbox_name' is the name given to the anatomical location within the Chest ImaGenome dataset, and is useful for lookups in other parts of the scene graph JSON; 'synsets' contain the UMLS CUI for the anatomical location concept; and the 'name' is the UMLS name for that CUI [4]. Note that CXRs are 2D images of a 3D structure so there are many overlying anatomical locations. A sample of 17 of the anatomical objects is plotted on a CXR as shown in Figure 2.

Each attribute dictionary, e.g., Supplementary (Section 3.3), aims to summarize all the CXR attribute descriptions for one anatomical location ('bbox_name'). This means, for a particular CXR anatomical location, all the sentences describing attributes related to it have been grouped into the 'phrases' field, where the order of sentences in the original report has been maintained. However, an anatomical location may not always be described or implied in the report. In that case, looking up dictionary['bbox_name'] will be False. The fields 'synsets' and 'name' are the same as in the objects' dictionaries, where they describe the UMLS CUI information for the anatomical location concept.

The 'attributes' field contains the relations between the anatomical location and the CXR attributes extracted from the respective sentences. Note that there can be multiple attributes extracted from each sentence. Therefore, the 'attributes' field is a list of lists. The 'attributes' in the lists follow the pattern of < categoryID | relation | label_name >, where 'categoryID' is the radiology semantic category the authors gave to the CXR concept in consultation with multiple radiologists, and relation is the NLP context relating the label_name to the anatomical location as an attribute. If the relation is 'no', then the 'label_name' is specifically negated in the sentence. If the relation is 'yes', then the 'label_name' is affirmed in the sentence. The order of the lists in the 'attribute_ids' field follow the lists in the 'attributes' field and map each 'label_name' to UMLS CUIs. Thus, the way the Chest ImaGenome dataset is formulated, one can interpret a statement such as the 'right lung' <has no> 'lung opacity' as true in the extracted radiology knowledge graph, whereby each node has been mapped to an externally recognized ontology.

The certainty of each relation in the CXR knowledge graph can be optionally further modified by the cues from the 'severity_cues' and 'temporal_cues' fields in each attribute dictionary. The severity cues can include 'hedge', 'mild', 'moderate' or 'severe', which are only assigned by co-occurrence at the sentence level. These extractions can benefit from future NLP improvement. Similarly, the temporal cues can modify the relation as either 'acute' or 'chronic' depending on clinical use cases.

The Chest ImaGenome categoryIDs can be used to differentiate the use case for different attributes:

• **anatomicalfinding** - findings of anatomies where there is some subjectivity in the grouping of the phrases used to extract the labels.

• **disease** - descriptions that are more diagnostic level and often require patient information outside the image and most subjective to the reading radiologist's inference/impression.

• **nlp** - normal / abnormal descriptions about different anatomical locations and can be subjective.

• **technicalassessment** - image quality issues affecting interpretation of CXR observations.

• **tubesandlines** - medical support devices where radiologists need to report any placement issues.

• **devices**: medical devices where placement issues are less relevant

• **texture** - these are only present in the 'texture_cues' field, we kept a set of highly non-specific attributes (e.g., opacity, lucency, interstitial, airspace) that tend to form the initial most objective descriptions about what is observed in the images by radiologists.

Finally, for comparison relationships, each dictionary has the format shown in Supplementary (Section 3.4). Each relationship dictionary describes the comparison relation(s) relevant for only one anatomical location ('bbox_name'). The 'relationship_id' uniquely identifies each comparison relationship between the object ('subject_id') on the current exam and the object ('object_id' for the same anatomical location) from the previous exam. The 'predicate' and 'synsets' are the UMLS CUIs for 'relationship_names', which is a list with usually one (but could be more) comparison relation type, which can be in ['comparison|yes|improved', 'comparison|yes|worsened', 'comparison|yes|no change']. The 'attributes' field records the attributes that are related to the anatomical location as per the sentence from the original report (kept in the 'phrase' field) that describes the comparison relationship.

CXR Scene Graphs Rendered in an Enriched RDF Format

Supplementary (Section 3.5): Radiology report sentences are fairly repetitive. Therefore, in the scene graph JSONS, one could see similar information described multiple times in different sentences for a study. In addition, in the MIMIC reports we worked with, each report could also have a preliminary read section (recorded by trainee radiologists - i.e., resident M.D.s) that comes before the final report section (approved by a fully trained and experienced radiologist). Thus, occasionally, the extraction from the sentences near the beginning of a CXR report can be different from the conclusion sentences

Metric	Sentence-level	Report-level
# of annotations	21,593	16,569
Precision	0.932	0.938
Recall	0.945	0.939
F1-score	0.939	0.939

Table 3: Object-attribute relations. Estimated inter-annotator (IA) agreement on 500 reports from first study: 0.984.

Metric	Sentence-level	Report-level
# of annotations	5,154 / 1,787	3,993 / 1,374
Precision	0.831 / 0.856	0.832 / 0.858
Recall	0.590 / 0.663	0.762 / 0.790
F1-score	0.690 / 0.747	0.796 / 0.823

Table 4: Object-object comparison relations (attribute-sensitive / attribute-blind). IA on 500 reports from second study: 0.962.

later in the report. To render the scene graphs easier for downstream utilization, we also provide post-processing utils (scenegraph_postprocessing.py) to roll the annotations up to the study level for each relation. This is done by taking the last relation extracted for each anatomical location and attribute combinations for a report. The processing utils can either render the scene graphs in a tabular format or represent the information in a simpler enriched RDF format, which we used to generate the graph visualizations in Figure 1.

Gold Standard Dataset Tables

We curated a manual gold standard evaluation dataset to measure the quality of the automatically derived annotations in the Chest ImaGenome dataset and for model benchmarking. Here we describe the three gold standard ground truth files in the "gold_dataset" directory. They are in tabular format for ease of comparison purposes.

• ***gold_attributes_relations_500pts_500studies1st.txt*** is the ground truth file which contains 21,594 object-to-attribute relations manually annotated for 3,042 sentences from the *first* CXR study for 500 unique patients. The notebook 'object-attribute-relation_evaluation.ipynb' explains in detail how we calculate the performance of object-to-attribute relation extraction.

• ***gold_comparison_relations_500pts_500studies2nd.txt*** is the ground truth file which contains 5,156 object-object (per attribute) comparison relations for 638 sentences from the *second* CXR study for the same 500 unique patients. The notebook 'object-object-comparison-relation_evaluation.ipynb' explains in detail how we calculate the performance for object-to-object-comparison relation extraction.

• The four ***bbox_coordinate_annotations*.csv*** files contain the manually annotated bounding box coordinates for the objects on the corresponding 1,000 unique CXR images. The notebook 'object-bbox-coordinates_evaluation.ipynb' calculates the bounding box object detection performance using these ground truth files.

• Lastly, ***final_merging_report_and_bbox_ground_truth.ipynb*** combines the manual text and anatomical bbox annotations as ***gold_object_attribute_with_coordinates.txt*** and ***gold_object_comparison_with_coordinates.txt***.

Additional supporting files for measuring the performance of the silver dataset against the gold standard are described in Supplementary (Section 5).

Dataset Evaluation

Table 3 ('analysis/generated via object-attribute-relation_evaluation.ipynb') reports the NLP pipeline's precision, recall and F1 scores for extracting the relationships between objects (anatomical locations) and CXR attributes (findings, diseases, technical assessment, etc) in the scene graphs. Since at their most granular level, the annotations are at the sentence-level, we report both the sentence-level and report-level results for 500 reports from the first exam of each patient. However, for most purposes, report-level annotations (the last annotation for each object-attribute relation for a study) are most suitable for downstream uses. The majority of the false positive results are due to failure to detect the laterality (i.e., left v.s. right) of attributes correctly as this information can often cross sentence boundaries, which is beyond the current NLP pipeline.

Table 4 (generated via 'analysis/object-object-comparison-relation_evaluation.ipynb') shows the NLP results for comparison relations (improved, worsened, no change) between various anatomical locations described for the current study as compared to the patient's previous study. The results are again shown at both sentence-level and report-level for 500 reports from the second exam of each patient. For the attribute-sensitive results, a relation is correct if it describes the correct comparison and

attribute for an object. Attribute-blind relations are correct as long as the object-to-object comparison relation is correct. Since comparison relations can cross both sentence and report boundaries, the performance from the current per sentence-based NLP pipeline is lower.

Lastly, Table 7 in Supplementary shows more detailed evaluation at the object-level (anatomical location). The F1 scores are calculated for relations extracted between objects and attributes from the 500 gold standard reports (first study), which is a breakdown of report-level results in Table 3 for the bounding boxes (Bboxes) shown. Using the $1,000$ CXR images in the gold standard dataset, we also calculated the intersection over union (IoU) between the automatically extracted Bboxes and the validated and corrected Bboxes (analysis/object-bbox-coordinates_evaluation.ipynb). Since we used an agree-or-correct annotation strategy for more efficient annotation, we also show the percentage of bounding boxes requiring manual correction in the gold dataset and the percentage missing in the final Chest ImaGenome dataset. Missing bounding boxes could be due to Bbox extraction failure or the anatomical location genuinely not being visible in the image (i.e., cut off or not in field of view), which is not uncommon for the costophrenic angles and apical zones. Per attribute level performance is available on the PhysioNet repository ('analysis/affirmed_attributes_eval4paper.csv').

Clinical Applications

There are numerous clinical topics that may be explored for a dataset that links anatomic structures with individual abnormalities and simultaneously provides comparison relation annotations for sequential images. Monitoring the progression of pathologies that are visualized through chest imaging is the most unexplored clinical application of this dataset. In the in-patient setting, diagnosis and monitoring of pneumonia are typically performed through comparisons of sequential CXR images from admission [23]. The same management principle may apply to the evaluation of the progression of other diseases, such as pneumothorax, pulmonary edema, acute respiratory distress syndrome, or congestive heart failure [6, 17, 37]. In the outpatient setting, surveillance of incidental pulmonary nodules, malignancies, tuberculosis, or interstitial lung disease is done through chest imaging in several-month intervals [15, 16, 25, 33]. Furthermore, the methodological concepts of this dataset could be extended to other modes of imaging, such as computed tomography (CT), and magnetic resonance (MR) imaging, etc, further expanding the potential clinical utility of this project.

Consistent dataset splits for performance reporting: For reproducibility, we include splits for train, valid and test sets in the "silver_dataset/splits" directory. Data split was performed at patient level. We also included a file (images_to_avoid.csv) with image IDs ('dicom_id') and 'study_id's for patients in the gold standard dataset, which should all be excluded from training and validation.

As described, Chest ImaGenome has been constructed with multiple possible downstream tasks in mind. Here, we showcase two example tasks that can have the most immediate clinical applications, (i) outputting both the location and the type of CXR attribute for an image (Example Task 2) and (ii) comparing whether a location has worsened or improved across sequential exams (Example Task 1). Clinically, the two chosen types of tasks are the two most important ones for radiologists to report when interpreting CXRs.

Example Task 1: Change between sequential CXR exams. CXRs are commonly repeatedly requested in the clinical workflow to assess for a myriad of attributes. Given a patient with sequential CXRs, the goal of this task is to automatically evaluate disease change over time based on two sequential CXR exams. We restricted the problem to a subset of the Chest ImaGenome dataset, i.e., to attributes related to congestive heart failure (CHF), as fluid management is one of the most routine clinical tasks for which CXRs can be ordered to guide the next steps (e.g., whether to give more intravenous fluid or give diuretics, etc). However, we note that users of this dataset can also explore comparison changes for other CXR attributes (e.g., pneumonia). Each CXR image is also associated with a bounding box that marks a localized area, e.g., "left lung" for specific anatomical finding (i.e., attribute), such as "pulmonary edema/hazy opacity", etc. In addition, the pair of CXR images is mapped to the comparison label that indicates whether the condition of the anatomical finding has improved or worsened. As a baseline example, we focus on change relations in the 'left lung' and 'right lung' objects that are related to the 'pulmonary edema/hazy opacity' and 'fluid overload/heart failure' attributes. The number of examples labeled in the training, validation and test data are $10,515$, $1,493$ and $2,987$, respectively.

We design a siamese architecture (Figure 11 in Supplementary 7) that first extracts the localized bounding box from each image and encodes the extracted image patches with a pre-trained ResNet101

Table 5: Anatomically localized CXR attribute detection (AUC scores). L1: Lung Opacity, L2: Pleural Effusion, L3: Atelectasis, L4: Enlarged Cardiac Silhouette, L5: Pulmonary Edema/Hazy Opacity, L6: Pneumothorax, L7: Consolidation, L8: Fluid Overload/Heart Failure, L9: Pneumonia.

Method	L1	L2	L3	L4	L5	L6	L7	L8	L9	**AVG**
Faster R-CNN	0.84	0.89	0.77	0.85	0.87	0.77	0.75	0.81	0.71	0.80
GlobalView	**0.91**	**0.94**	0.86	0.92	0.92	**0.93**	0.86	0.87	0.84	0.89
CheXGCN	0.86	0.90	**0.91**	**0.94**	**0.95**	0.75	**0.89**	**0.98**	**0.88**	**0.90**

autoencoder, denoted that is trained on several medical imaging datasets, e.g., NIH, CheXpert, and MIMIC datasets, etc. [20, 22, 42]. The autoencoder image representations are concatenated and passed through a dense layer with 128 neurons and ReLU activations, and a final classification layer.

We train for 300 epochs with cross-entropy, stochastic gradient descent, $1e-3$ learning rate, 0.1 gradient clipping and 32 batch size. We freeze the autoencoder weights and finetune the two last dense layers. On this challenging task of predicting change in localized anatomical findings between two sequential exams, we achieve an accuracy of 75.3%.

Example Task 2: Localization of CXR attributes. Knowing the anatomical location of non-specific findings/attributes on CXR images can help with narrowing down possible disease diagnoses and guide the next steps in requesting more specific imaging exams or treatment. To this end, we train a Faster R-CNN model [35] to learn 18 anatomical locations within the dataset. We extract the 1024 dimension convolution feature vector of each anatomical region. We re-implement the state-of-the-art CheXGCN model [7] to learn the dependencies between attributes within the Chest X-ray. Similar to the work done by CheXGCN we model the correlation of the CXR attributes using a conditional probability (see Figure 12 in Supplementary). We compare the results of the model with two baseline models, a Faster R-CNN model followed by a linear model without the GCN, and a Densenet model [19] without the Faster R-CNN to evaluate the effectiveness of the localized models. We focus on 9 common CXR attributes, which include lung opacity, pleural effusion, atelectasis, enlarged cardiac silhouette, pulmonary edema/hazy opacity, pneumothorax, consolidation, fluid overload/heart failure, pneumonia. The results of the experiments are shown in Table 5 and the labels are ordered according to the attribute list above.

Dataset Limitations: The Chest ImaGenome dataset came from only one U.S. hospital sourcem, is automatically generated and is limited by the performance of the NLP and the Bbox extraction pipelines. Cross-sentence relations, as well as other longer dependencies displayed in the text are not captured, and text and image annotations are performed separately. Furthermore, we cannot assume that all the clinically relevant CXR attributes are always described on every exam by the reporting radiologists. In fact, we have observed many implied object-attribute relation descriptions that are documented only in the form of comparisons (e.g., no change from previous) in short CXR reports. As such, even with perfect NLP extraction of object and attribute relations from individual reports, there would be missing information in the report knowledge graph constructed for some images. These technical areas are worth improving on in future research with more powerful NLP, image processing techniques and other graph-based techniques. Addressing missing relations will certainly improve this dataset too. Regardless, version 1.0.0 of the Chest ImaGenome dataset serves as a pioneering vision for a richer radiology imaging dataset.

Conclusions and Future Work

In this paper, we release Chest ImaGenome, the first VisualGenome-style Chest X-Ray (CXR) dataset to facilitate clinical reasoning research. Chest ImaGenome includes $1,256$ relation combinations between 29 CXR anatomical locations and their attributes, structured as a scene graph per image, and over $670,000$ localized comparison relations across sequential exams. We invite the research community to expand on our work. To this end, we describe the potential utility of this dataset in clinical applications and present experiments on two example tasks. In the near future, we plan to extend to more data sources and incorporate a VQA component.

Acknowledgements

This work was supported by the Rensselaer-IBM AI Research Collaboration, part of the IBM AI Horizons Network, and the IBM-MIT Critical Data Collaboration.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1955–1960, 2016.

[2] Nkechinyere N Agu, Joy T Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. Anaxnet: Anatomy aware multi-label finding classification in chest x-ray. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

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

[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[6] Luciano Cardinale, Adriano Massimiliano Priola, Federica Moretti, and Giovanni Volpicelli. Effectiveness of chest radiography, lung ultrasound and thoracic computed tomography in the diagnosis of congestive heart failure. *World journal of radiology*, 6(6):230, 2014.

[7] Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE journal of Biomedical and Health Informatics*, 24(8):2292–2302, 2020.

[8] Anni Coden, Daniel Gruhl, Neal Lewis, Michael Tanenblatt, and Joe Terdiman. Spot the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In *Proceedins for the 2nd IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 33–39. IEEE, 2012.

[9] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335, 2017.

[10] Surabhi Datta and Kirk Roberts. A dataset of chest x-ray reports annotated with spatial role labeling annotations. *Data in Brief*, 32:106056, 2020.

[11] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5503–5512, 2017.

[12] 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, 2016.

[13] Ross W Filice, Anouk Stein, Carol C Wu, Veronica A Arteaga, Stephen Borstelmann, Ramya Gaddikeri, Maya Galperin-Aizenberg, Ritu R Gill, Myrna C Godoy, Stephen B Hobbs, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset. *Journal of Digital Imaging*, 33(2):490–496, 2020.

[14] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5630–5639, 2017.

[15] Michael K Gould, Jessica Donington, William R Lynch, Peter J Mazzone, David E Midthun, David P Naidich, and Renda Soylemez Wiener. Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5):e93S–e120S, 2013.

[16] David M Hansell, Jonathan G Goldin, Talmadge E King Jr, David A Lynch, Luca Richeldi, and Athol U Wells. Ct staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the fleischner society. *The Lancet Respiratory Medicine*, 3(6):483–496, 2015.

[17] M Henry, T Arnold, and J Harvey. Bts guidelines for the management of spontaneous pneumothorax. *Thorax*, 58(Suppl 2):ii39, 2003.

[18] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. *Zenodo*, 2020.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[20] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 590–597, 2019.

[21] 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.

[22] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, pages 1–8, 2019.

[23] Andre C Kalil, Mark L Metersky, Michael Klompas, John Muscedere, Daniel A Sweeney, Lucy B Palmer, Lena M Napolitano, Naomi P O'Grady, John G Bartlett, Jordi Carratalà, et al. Management of adults with hospital-acquired and ventilator-associated pneumonia: 2016 clinical practice guidelines by the infectious diseases society of america and the american thoracic society. *Clinical Infectious Diseases*, 63(5):e61–e111, 2016.

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[25] Hyun Jung Koo, Chang-Min Choi, Sojung Park, Han Na Lee, Dong Kyu Oh, Won-Jun Ji, Seulgi Kim, and Mi Young Kim. Chest radiography surveillance for lung cancer: Results from a national health insurance database in south korea. *Lung Cancer*, 128:120–126, 2019.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[27] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1537–1547, 2018.

[28] Matthew D Li, Nishanth Thumbavanam Arun, Mishka Gidwani, Ken Chang, Francis Deng, Brent P Little, Dexter P Mendoza, Min Lang, Susanna I Lee, Aileen O'Shea, et al. Automated assessment and tracking of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2(4):e200079, 2020.

[29] Matthew D Li, Ken Chang, Ben Bearce, Connie Y Chang, Ambrose J Huang, J Peter Campbell, James M Brown, Praveer Singh, Katharina V Hoebel, Deniz Erdoğmuş, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ digital medicine*, 3(1):1–9, 2020.

[30] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1261–1270, 2017.

[31] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[32] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[33] Payam Nahid, Susan E Dorman, Narges Alipanah, Pennan M Barry, Jan L Brozek, Adithya Cattamanchi, Lelia H Chaisson, Richard E Chaisson, Charles L Daley, Malgosia Grzemska, et al. Official american thoracic society/centers for disease control and prevention/infectious diseases society of america clinical practice guidelines: treatment of drug-susceptible tuberculosis. *Clinical Infectious Diseases*, 63(7):e147–e195, 2016.

[34] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[36] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045, 2018.

[37] GD Rubenfeld, T Thompson, ND Ferguson, E Caldwell, E Fan, L Camporota, and AS Slutsky. Acute respiratory distress syndrome. the berlin definition. *JAMA*, 307(23):2526–2533, 2012.

[38] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[39] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020.

[40] PLCO Project Team, John K Gohagan, Philip C Prorok, Richard B Hayes, and Barnett-S Kramer. The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. *Controlled Clinical Trials*, 21(6):251S–272S, 2000.

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[42] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[43] 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 (CVPR)*, pages 9049–9058, 2018.

[44] Joy Wu, Yaniv Gur, Alexandros Karargyris, Ali Bin Syed, Orest Boyko, Mehdi Moradi, and Tanveer Syeda-Mahmood. Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 799–803. IEEE, 2020.

[45] Joy T Wu, Ali Syed, Hassan Ahmad, et al. Ai accelerated human-in-the-loop structuring of radiology reports. In *Americal Medical Informatics Association (AMIA) Annual Symposium*, 2020.

[46] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017.

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015.

[48] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018.

[49] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12910–12917, 2020.

[50] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, 2018.