Uncovering Knowledge Gaps in Radiology Report Generation Models through Knowledge Graphs

Xiaoman Zhang, Julián N. Acosta, Hong-Yu Zhou, Pranav Rajpurkar

Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

https://github.com/rajpurkarlab/ReXKG

Abstract

Recent advancements in artificial intelligence have significantly improved the automatic generation of radiology reports. However, existing evaluation methods fail to reveal the models' understanding of radiological images and their capacity to achieve human-level granularity in descriptions. To bridge this gap, we introduce a system, named ReXKG, which extracts structured information from processed reports to construct a comprehensive radiology knowledge graph. We then propose three metrics to evaluate the similarity of nodes (ReXKG-NSC), distribution of edges (ReXKG-AMS), and coverage of subgraphs (ReXKG-SCS) across various knowledge graphs. We conduct an in-depth comparative analysis of AI-generated and human-written radiology reports, assessing the performance of both specialist and generalist models. Our study provides a deeper understanding of the capabilities and limitations of current AI models in radiology report generation, offering valuable insights for improving model performance and clinical applicability.

Introduction

Artificial Intelligence (AI) models have recently achieved remarkable success in interpreting medical images (Rajpurkar and Lungren 2023; Rajpurkar et al. 2022). Among them, radiology report generation stands out as a crucial task in medical imaging, providing essential information for further diagnosis and treatment planning (Liu, Tian, and Song 2023; Reale-Nosei et al. 2024). Its significance has led to a surge in research focused on developing AI models capable of generating these reports (Zhang et al. 2020; Liu et al. 2024). However, in-depth understanding radiology report generation models' performance is a challenging yet important task for real clinical usage.

Various automated evaluation metrics have been proposed specifically for report generation, such as RadCliQ (Yu et al. 2023), FineRadScore (Huang et al. 2024), RaTEScore (Zhao et al. 2024) and GREEN (Ostmeier et al. 2024), *etc*. These metrics have gradually approached the quality of radiologists' evaluations. Yet, most existing metrics rely on report-to-report comparisons, which fail to fully capture a model's holistic understanding of radiological images or its capacity to match the descriptive granularity used by humans. For example, when a doctor mentions "edema" in a report, they may use nuanced modifiers such as "moderate",

"mild", "unchanged", "decreased", or "stable" to convey precise details. In contrast, a model might not capture this level of detail or variation in terminology. It is essential to develop evaluation methods considering the comprehensiveness of medical terminology understanding. These insights can guide the improvement of report generation models, ensuring they are better aligned with the professional descriptions used by radiologists.

In this paper, we target assessing AI models from a different perspective by focusing on the radiological knowledge learned by the model. To accomplish this, we introduce a system named **ReXKG**, designed to extract structured information from processed reports and construct a comprehensive radiology knowledge graph. As shown in Figure 1, this graph will capture relationships between anatomical structures, pathologies, imaging findings, medical devices, and procedures, creating a rich, queryable representation of radiological knowledge. We propose three novel metrics: ReXKG-NSC for assessing node similarity, ReXKG-AMS for evaluating edge distribution, and ReXKG-SCS for measuring subgraph coverage across knowledge graphs. These metrics allow for a global score comparison between models and against human radiologists, providing a comprehensive understanding of the model's performance.

Based on the knowledge graph and proposed metrics, we conduct a comprehensive analysis of both specialist and generalist report generation models, exploring the following questions and summarizing the main conclusions for each:

Q1: Coverage of Entities. How well do the generated reports cover essential entities such as anatomy and disorders? Generalist models demonstrate broader coverage, capturing nearly 80% of essential entities, yet they still fall short of matching the depth of radiologist-written reports, particularly in detailing medical devices.

Q2: Coverage of Relationships Between Entities. How comprehensively do the AI reports describe connections between different medical findings and their descriptions? All AI models show significant gaps compared to radiologist-written reports in capturing relationships between different entities, with MedVersa leading, achieving nearly 80% coverage of the top 10% subgraphs.

Q3: Coverage of Concepts or Descriptors. How detailed and comprehensive are the descriptions of disorders and anatomical features? AI models tend to overfit specific con-
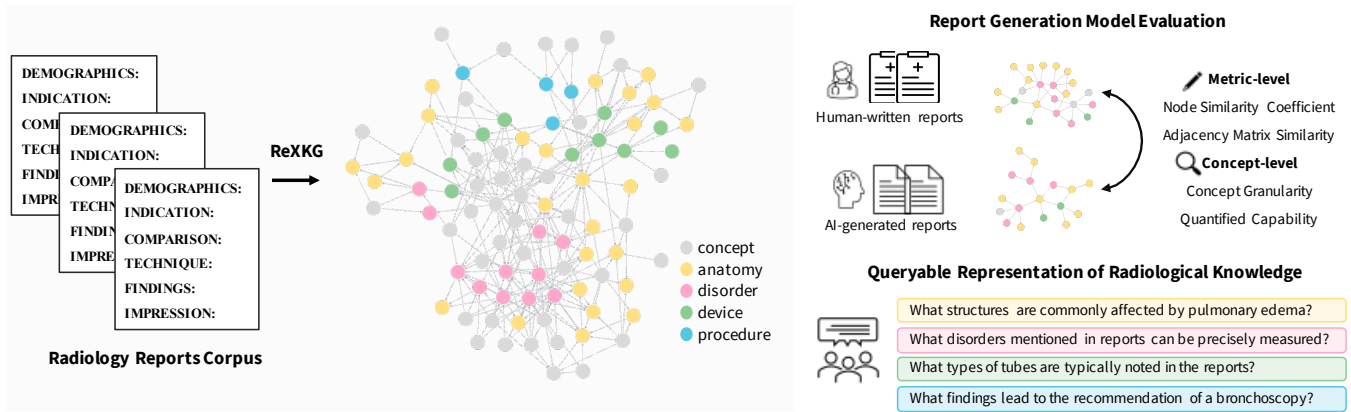
Figure 1: An illustration of **Learning from Knowledge Graph**.

cepts that appear frequently in the training data, resulting in less detailed and occasionally hallucinated descriptions.

Q4: Quantitative Measurements Coverage. How frequently does the model provide quantified measurements of disorders? AI model's behavior in providing size descriptions correlates strongly with the frequency of size descriptions for specific disorders in the training data.

Q5: Specialist vs. Generalist Models. What are the performance differences between specialist and generalist models? Generalist models, trained on multiple modalities of data, demonstrate significantly enhanced radiology knowledge compared to specialist models. This suggests that exposure to a broader range of medical data and tasks contributes to a more comprehensive and accurate representation of radiological concepts and relationships.

Knowledge Graph Construction

In this section, we present our system (**ReXKG**) for constructing a comprehensive knowledge graph from a large corpus of radiology reports, shown in Figure 2. We first define an information extraction schema tailored to the radiology domain, then once the entities and relationships are extracted, we proceed with the node construction pipeline to ensure data consistency and integrity. Finally, we integrate the information into the graph structure.

Information Extraction Schema

Definition. We define an entity as a continuous span of text that can include one or more adjacent words. Entities in our schema are categorized into six types as listed.

- **Anatomy**: anatomical structures within the body.
- **Disorder**: any abnormal findings or diseases identified within radiology reports.
- **Concept**: descriptors used to modify other entities, for example, "acute", "severe", and "increasing".
- **Device**: any instrument or apparatus used for medical purposes, for example, "tube", "clip", "wire".
- **Procedure**: medical procedures used to diagnose, measure, monitor, or treat conditions, such as "sternotomy".

- **Size**: measurements of disorders or anatomical structures, for example, "3-mm".

We define a relation as a directed edge between two entities. Following the previous work (Jain et al. 2021a), our schema uses three relations as listed.

- **Suggestive of**: source entity (e.g., findings) may suggest the presence of the target entity (e.g., a disease).
- **Located at**: source entity is located at the target entity.
- **Modify**: source entity modifies or provides additional information about the target entity.

Entity and Relation Extraction. Given a set of radiology reports, we first annotate a subset using GPT-4 (Achiam et al. 2023) to generate labeled entities and relations. The prompts used for annotation are provided in the appendix. Based on the annotated data, we train the model using the Princeton University Relation Extraction system (PURE) architecture (Zhong and Chen 2021) to do Named Entity Recognition (NER). This architecture employs a pipeline approach, decomposing the tasks of entity recognition and relation extraction into separate subtasks. Once the model is trained, we apply it to the entire dataset to perform inference, extracting all relevant entities and relations.

Nodes Construction

Following entity extraction, we employ a series of steps to remove noise, merge synonyms, and link entities to the Unified Medical Language System (UMLS) (Bodenreider 2004). First, we determine the entity type of each extracted entity based on the most frequently predicted type by the NER model to ensure consistency and accuracy. Next, we utilize ScispaCy (Neumann et al. 2019) to retrieve UMLS attributes for each entity, such as Concept Unique Identifiers (CUI), Type Unique Identifiers (TUI), definitions, and aliases. Entities that cannot be mapped to a UMLS item are retained for further processing. For entities identified as aliases of a specific term in UMLS, we normalize these entities by merging them into a single concept. For instance, entities such as "pulmonary" and "lung" are normalized to their corresponding CUI C0024109. Additionally, to ensure

a. Information Extraction System

Unchanged CONCEPT position CONCEPT of the left CONCEPT upper CONCEPT extremity ANATOMY PICC line DEVICE-PRESENT. There are persistent CONCEPT small CONCEPT bilateral CONCEPT effusions DISORDER-PRESENT. Absence CONCEPT of the right CONCEPT breast ANATOMY shadow CONCEPT compatible with prior CONCEPT mastectomy PROCEDURE.

Entity Extraction

(Unchanged, position, modify) (left, extremity, modify)
(upper, extremity, modify) (PICC line, extremity, located at)
...... *(bilateral, effusions, modify)* ......
(shadow, breast, located at) (prior, mastectomy, modify)

Relation Extraction

b. Nodes Construction
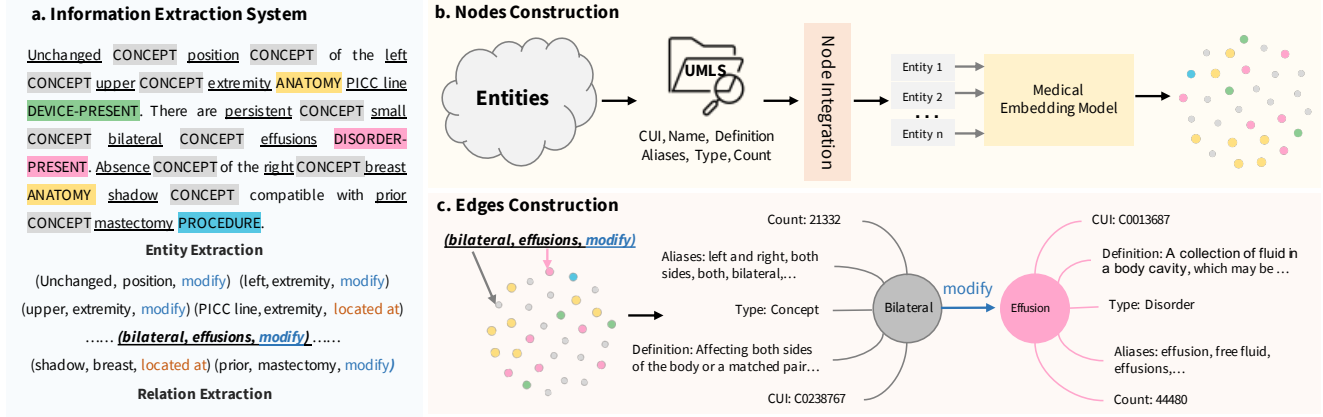
c. Edges Construction

Figure 2: Overview of the proposed knowledge graph construction system **ReXKG**. (a) The information extraction system for entity and relation extraction. (b) The node construction pipeline. (c) Illustration of edge construction.

the compactness and unambiguity of nodes, for the multi-word entities, if all individual words of such an entity are predicted as separate nodes, the combined multi-word entity is not included as a node. The detailed algorithm is provided in the appendix. Finally, we leverage medical language models to merge entities based on semantic similarity. Entities with an embedding similarity higher than a defined threshold are combined. This step enhances the graph's coherence by aggregating semantically similar concepts into single nodes.

Edges Construction

Initially, all relations are extracted from the dataset as triplets (source entity, target entity, relation). We merge different triplets with the same source and target entities based on node aliases. When two nodes are linked by multiple relation types, we retain the relation type most frequently predicted by the model. Finally, we filter the relations by ignoring triplets with a count less than C, a hyperparameter ensuring the reliability of the connections within the graph.

Knowledge Graph Evaluation Metrics

To evaluate knowledge graphs obtained from different models, we introduce three metrics that assess node similarity, edge distribution similarity, and subgraph coverage: **ReXKG-NSC** (Node Similarity Coefficient), **ReXKG-AMS** (Adjacency Matrix Similarity), and **ReXKG-SCS** (Subgraph Coverage Score). In the following, we will first provide a preliminary definition of the knowledge graph and then detail the calculation methods for these metrics.

Preliminary Definition

Assume we have a knowledge graph with N nodes and M edges. The set of nodes is denoted as $V = \{v_1, v_2, \ldots, v_N\}$. The weights of the nodes are represented as $W_V = \{w_{v_1}, w_{v_2}, \ldots, w_{v_N}\}$, where where w_{v_i} corresponds to the frequency of node v_i in the data. The set of edges is denoted as $E = \{e_1, e_2, \ldots, e_M\}$, where each edge e_m connects a pair of nodes (v_i, v_j). The weights of the edges are represented as $W_E = \{w_{e_1}, w_{e_2}, \ldots, w_{e_M}\}$, where $w_{e_m} =$

count(e_m). Then, the adjacency matrix is defined as A, with $A_{ij} = w_{e_{ij}}$, representing the weight of the edge between nodes v_i and v_j.

KG Node Similarity Coefficient

Let KG-GT represent the knowledge graph built from the ground truth reports, consisting of N nodes. Similarly, let KG-Pred represent the knowledge graph built from the generated reports, consisting of P nodes. For each node v_i in KG-GT, we identify the most similar node in KG-Pred, assigning a similarity score s_i based on calculations from a medical language model. The overall node similarity metric is then calculated as the average of these similarity scores across all nodes in KG-GT. This can be expressed as:

$$\texttt{KG-NSC} = \frac{1}{N} \sum_{i=1}^{N} s_i. \quad (1)$$

KG Adjacency Matrix Similarity

For each node v_i in KG-GT, we identify the most similar node in KG-Pred. This allows us to map all edges in KG-Pred using the nodes from KG-GT, resulting in the creation of two adjacency matrices, A_{Pred} and A_{GT}, both of the same size. Where A_{ij} represents the weight of the edge between nodes i and j. We use the Pearson correlation coefficient metrics to evaluate the coverage of relations in generated reports compared to the ground truth. The row weight w_{r_i} is used as the weight, and the Pearson correlation coefficient as the value. Here, for a given row i, the row weight is defined as $w_{r_i} = (\sum_j A_{ij})/(\sum_i \sum_j A_{ij})$, where A_{ij} represents the element at row i, column j of the adjacency matrix. Thus, the adjacency matrix similarity can be expressed as:

$$\texttt{KG-AMS} = \frac{\sum_{i=1}^{N} \left(w_{r_i} \cdot \texttt{corr}(A_{Pred,i}, A_{GT,i})\right)}{\sum_{i=1}^{N} w_{r_i}}, \quad (2)$$

where $\texttt{corr}(A_{Pred,i}, A_{GT,i})$ is the Pearson correlation coefficient between the i-th rows of A_{Pred} and A_{GT}, and w_{r_i} is the weight of all edges associated with the i-th row.

KG Subgraph Coverage Score

Let $\mathcal{S}$ be the set of all connected subgraphs in KG-GT up to a size of k nodes. We quantify a model's ability to represent important subgraphs from KG-GT within KG-Pred, which can be expressed as:

$$\text{KG-SCS} = \frac{\sum_{i=1}^{K} I(S_i) \cdot P(S_i)}{\sum_{i=1}^{K} I(S_i)}, \tag{3}$$

where K is the number of top important subgraphs considered. $I(S_i)$ is the importance score of each subgraph S_i in KG-GT and $P(S_i)$ is the presence score in KG-Pred. Please refer to the appendix for detailed definitions.

Experiments

In this section, we present the dataset and models used in our analysis of AI-generated reports. Given the current limitations in model capabilities, with few models available for generating CT/MRI reports, our study primarily focuses on chest X-ray report analysis. However, the proposed ReXKG is versatile and applicable across various modalities and anatomical regions, as demonstrated in the appendix.

Datasets

CheXpert Plus: CheXpert Plus (Chambon et al. 2024) is a dataset that pairs text and images, featuring 223,228 unique pairs of radiology reports and chest X-rays from 187,711 studies and 64,725 patients. Each patient may be linked to multiple studies, and each study may include several images.

MIMIC CXR: MIMIC-CXR (Johnson et al. 2019) is a large publicly available dataset of chest X-rays with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center.

Experiments Settings

To ensure a comprehensive analysis, we randomly split studies from CheXpert Plus into two parts: CheXpert Plus I (24,086 studies) and CheXpert Plus II (24,085 studies). Additionally, we randomly select a subset from MIMIC-CXR with 24,085 studies for comparison. We designate CheXpert Plus I as the benchmark for our study. This subset serves as the ground truth, upon which all model evaluations are conducted, inference tasks performed, and knowledge graphs constructed. Similarly, we can set CheXpert Plus II as the benchmark, with results provided in the appendix. The knowledge graphs for comparison can be categorized into two groups based on the data source.

Intra-Dataset Reports: Intra-Dataset Reports are knowledge graphs built from real clinical datasets across different studies or centers. We use CheXpert Plus II and the selected MIMIC-CXR subset, which represent radiologist-written reports from various studies and centers, as benchmark baselines for comparison with AI-generated reports.

Extra-Dataset Reports: Extra-Dataset Reports are knowledge graphs constructed from AI-generated reports. To comprehensively evaluate AI performance, we assess various report generation models, including specialist models such as CvT2DistilGPT2 (Nicolson et al. 2023), RGRG (Tanida et al. 2023), and Swinv2-MIMIC (Chambon et al. 2024), as well as generalist models like CheXagent (Chen et al. 2024), RadFM (Wu et al. 2023), and MedVersa (Zhou et al. 2024). Here, specialist models are defined as those trained exclusively on chest X-ray report generation, whereas generalist models are large-scale models trained on various tasks. Details of these models can be found in the appendix.

Implementation Details

For the Information Extraction Schema, we follow the approach described in (Jain et al. 2021a), utilizing the PURE framework (Zhong and Chen 2021), which employs a pretrained BERT model to obtain contextualized representations. These representations are then fed into a feedforward network to predict the probability distribution of entities, which subsequently serves as input for the relation model. The learning rate is set to 2e-5 during training. We use Med-CPT (Jin et al. 2023) as the default medical language model for entity merging, with a merging threshold of 0.95. The threshold C for edge construction is set to 5. The number of nodes in each subgraph is set to $k = 2$, and the number of important subgraphs, K, is defined as 10% of the total subgraphs in KG-GT. For report generation inference, we use the code and checkpoints provided by the respective baseline models, focusing on the generation of the findings section. All experiments are conducted on an NVIDIA A100 GPU.

Results

In this section, we present a comprehensive analysis of knowledge graphs generated from both intra-dataset reports (radiologist-written) and extra-dataset reports (AI-generated). Using CheXpert Plus I as our benchmark, we hypothesize that the knowledge graph generated from CheXpert Plus II will display similar nodes, edges, and distribution characteristics. Such similarity would validate the consistency of our findings and underscore the reliability and quality of our proposed methods for constructing knowledge graphs. Our analysis is structured around key questions that probe different aspects of report generation, from entity coverage to relationship comprehension, providing a multi-faceted view of current AI models' capabilities.

Q1: Coverage of Entities

First, we explore the question: **How well do the AI-generated reports cover essential entities such as concepts, anatomy, disorders, devices, and procedures?**

As shown in Table A1, We compare the KG-NSC between CheXpert Plus I with other datasets and various report generation models. CheXpert Plus II and MIMIC-CXR, representing radiologist-written reports with similar and differing distributions of ground truth, exhibit high similarity across all entity types, with overall scores of 0.970 and 0.928. This high similarity demonstrates the reliability of the proposed metric and sets a high benchmark for AI models to match. Among AI models, generalist models, particularly RadFM and MedVersa, exhibit broader coverage of essential entities

Type	Models	KG-NSC						KG-AMS				KG-SCS
		Ana.	Dis.	Con.	Dev.	Pro.	All	Dis.Ana.	Dev.Ana.	Dis.Dis.	All	k=2
Intra-Dataset	CheXpert Plus II	0.974	0.967	0.970	0.958	0.977	0.970	0.966	0.981	0.988	0.971	0.981
	MIMIC-CXR	0.930	0.948	0.930	0.865	0.929	0.928	0.841	0.786	0.858	0.819	0.950
Specialist	CvT2DistilGPT2 (Nicolson et al. 2023)	0.781	0.760	0.786	0.730	0.809	0.779	0.776	0.841	0.752	0.624	0.696
	RGRG (Tanida et al. 2023)	0.657	0.627	0.624	0.589	0.577	0.626	0.681	0.680	0.642	0.579	0.538
	Swinv2-MIMIC (Chambon et al. 2024)	0.772	0.773	0.772	0.742	0.782	0.777	0.719	0.814	0.821	0.646	0.648
Generalist	CheXagent (Chen et al. 2024)	0.720	0.698	0.707	0.675	0.716	0.707	0.856	**0.883**	0.567	0.710	0.588
	RadFM (Wu et al. 2023)	**0.817**	0.829	0.796	0.732	0.777	0.800	0.725	0.695	0.538	0.601	0.733
	MedVersa (Zhou et al. 2024)	0.807	**0.830**	**0.801**	**0.754**	**0.818**	**0.804**	**0.859**	0.843	**0.894**	**0.748**	**0.806**

Table 1: Knowledge graph comparison between CheXpert Plus I and Intra-Dataset or Extra-Dataset Reports. KG-NSC, KG-AMS, and KG-SCS scores are reported. The best results are highlighted in boldface.

compared to specialist models. This superior performance likely stems from their training on more diverse and large-scale datasets, enabling these models to generalize better and capture a wider range of medical entities.

When examining the results for each entity type, there is a noticeable gap in medical devices across all models. This discrepancy may be attributed to the primary factor that models are exclusively trained on the MIMIC-CXR dataset, thus the models' predictions align more closely with MIMIC-CXR's distribution. However, there are inherent distribution differences between the CheXpert Plus and MIMIC-CXR datasets. CheXpert Plus includes some rare devices, such as the "Impella", which is mentioned only 15 times in the entire CheXpert Plus dataset. Additionally, varied terminology is used to describe the type of devices, such as "keofeed" for "tubes".

Q2: Coverage of Relationships Between Entities

Next, we investigate **How comprehensive is the coverage of relationships between entities?**
To evaluate the comprehensiveness of AI-generated reports in capturing relationships between entities, we employed the KG-AMS and KG-SCS metrics. Table A1 details the correlation between specific types of relationships: disorders with anatomy, devices with anatomy, and relationships between disorders. MedVersa leads in the KG-AMS metric across most categories, particularly excelling in disorder-disorder and overall relationships. CheXagent, on the other hand, stands out in device-anatomy relationships, while RadFM shows balanced performance across various types of entity relationships. Despite these performances, there remains a significant gap compared to radiologist-written reports, highlighting areas for further improvement. The KG-SCS metric (with k=2) offers additional insights into how well models capture important subgraphs or patterns within the knowledge graph. MedVersa covers 80.6% of the important subgraphs, while RadFM covers over 73%, indicating that while these models perform well, there is still room for enhancement in capturing complex relationships.

Q3: Comprehensiveness of Concepts

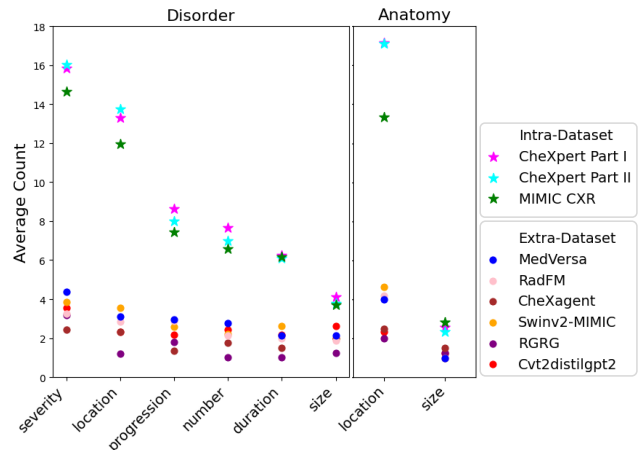We further access the quality of content generated by AI models with the question: **How detailed and comprehen-**

Figure 3: Average count of concept entities used to modify disorders and anatomy across different models.

sive are the descriptions of disorders and anatomical regions provided by the AI models?

This question is critical for applying AI models in clinical scenarios, where the ability to describe and differentiate the severity of diseases can directly impact diagnosis and treatment planning. To assess the depth and comprehensiveness with which disorders and anatomical regions are described, we utilize GPT-4 to classify all concept nodes within our knowledge graphs. These concepts are categorized into the following:

- **Severity**: Describes how intense or severe the symptoms are, such as mild, moderate, or severe.
- **Location**: Specifies where on or in the body the disorder manifests, such as left, right, bilateral, upper, lower, or specific organs or systems involved.
- **Duration**: Refers to how long the disorder or its symptoms have been present. (acute, chronic, transient)
- **Progression**: Indicates how the disorder changes over time, including progressive, stable, and regressive.
- **Size**: Relevant for physical abnormalities or tumors, indicating how large an affected area or lesion is.
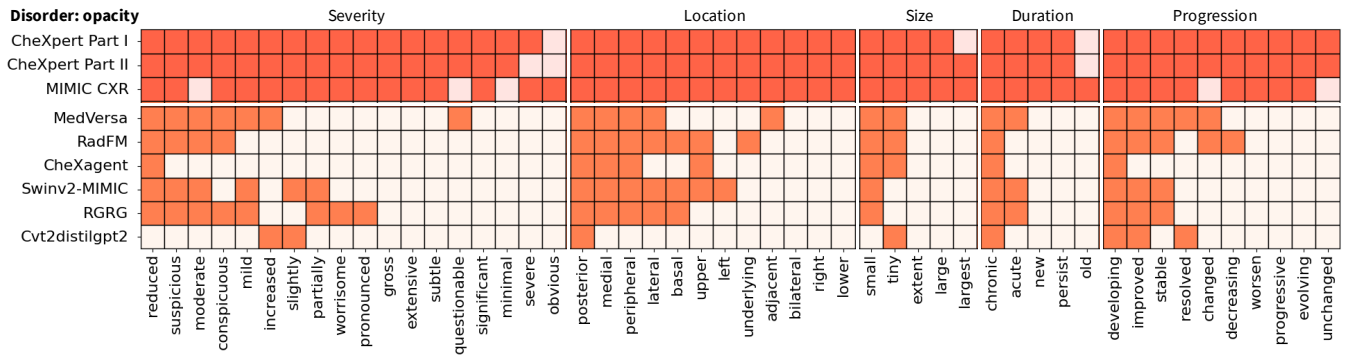
Figure 4: Detailed results of model predictions for given concepts related to specific disorders. Dark orange indicates the model predicts the relationship, while light orange indicates not.

- **Number**: Describes how many lesions or abnormalities are present, such as single, multiple, or widespread.

Our analysis, depicted in Figure 3, shows that Intra-Dataset groups exhibit the highest similarity, with nearly identical counts for all category concepts used to modify disorders and anatomy. In contrast, AI models tend to underperform, especially in categories like "severity" and "location". Models often describe "location" for anatomy and "severity" for disorders, such as specifying "left lung" or "mild edema", but the range of terms they use for modification is limited. Moreover, since all models perform inference without considering prior studies, concepts related to progression such as "unchanged" or "improved" may result from hallucinations This issue arises partly because the training data often lack comprehensive, longitudinal information that accurately captures patient progression. Additionally, some model training processes do not take into account the patient history or the continuity of patient data across multiple studies.

To gain a more detailed understanding, we selected several high-frequency disorders and the commonly used concepts to modify these disorders. One example is shown in Figure 4, the Intra-Dataset Reports's results exhibit complete coverage. In contrast, models tend to use concepts like "moderate" and "mild" but do not use terms "severe" or "subtle" for "opacity". We provide comprehensive detailed results in the appendix, from which we can observe that for some disorders, such as "consolidation", most models do not provide severity descriptions. We also provide a barplot in the appendix showing the frequency of those concepts in MIMIC-CXR training set, an interesting observation is that the model's predictions are not linearly related to the frequency of appearance in the MIMIC-CXR training set. Instead, the model tends to overfit a specific synonym within a set of related concepts, and the selected concept varies for different disorders.

Q4: Quantified Measurement

We then address the issue of quantification in the reports: **How frequently does the model provide quantified measurements of disorders and anatomical regions?**
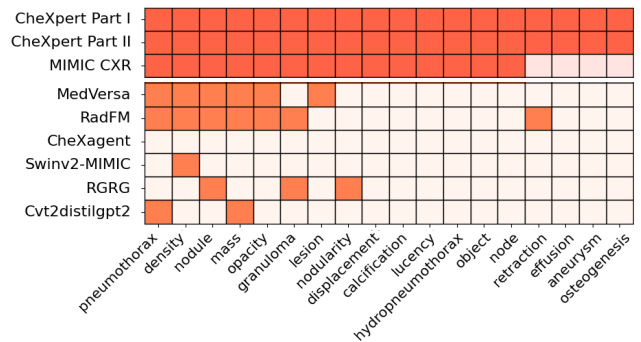

Figure 5: Detailed results of whether the model predicts specific size measurements for given disorders. Dark orange indicates the model predicts, while light indicates not.

This information is crucial for the deep analysis of images. For instance, disorders that consistently include size descriptions like "3mm" in the report might require the development of precise segmentation targets. On the other hand, some disorders that cannot be measured may only need bounding boxes during labeling. Based on the knowledge graph, AI research can easily identify which disorders can and should be segmented, thereby further promoting research on grounded report generation.

As shown in Figure 5, we provide an overview of whether the models give detailed measurement descriptions for the target disorders. Both CheXpert Part I and CheXpert Part II consistently provide detailed descriptions for all target disorders, which highlights the real clinical requirements. However, most AI models show limited coverage, often failing to provide detailed descriptions for many conditions like calcification and effusion. Relatively speaking, generalist models like RadFM and MedVersa cover a broader range of disorders. It is notable that CheXagent does not predict any size measurements for disorders but consistently provides size descriptions for devices such as tubes and lines. We also provide the frequency of size descriptions for specific disorders in MIMIC-CXR training data in the appendix, as shown, the model's behavior in providing size descriptions correlates strongly with the frequency.

Type	Models	BLEU	BERT	Semb	RadG	RadC
Specialist	CvT2DistilGPT2	0.123	0.262	0.286	0.119	1.585
	RGRG	**0.141**	**0.304**	0.257	**0.127**	1.533
	Swinv2-MIMIC	0.129	0.286	0.284	0.123	1.543
Generalist	CheXagent	0.102	0.299	0.294	0.124	1.510
	RadFM	0.091	0.259	0.202	0.083	1.718
	MedVersa	0.116	0.300	**0.315**	**0.127**	**1.483**

Table 2: Comparisons of both specialist and generalist models on CheXpert Plus. Metrics include BLEU, BERTScore (BERT), SembScore (Semb), RadGraph F1 (RadG), and RadCliQ-v1 (RadC).

Q5: Specialist vs. Generalist Models

Finally, we compare different types of AI models by asking: **What are the differences in performance between specialist models and generalist models?**

We summarize the score of different metrics on different models' predictions on the training set of CheXpert Plus finding sections. Note that none of the models were trained using CheXpert Plus. First, we observe that there is not a significant gap between the report-vs-report performance scores of specialist models and generalist models. This suggests that specialist models can perform well on specialist tasks. However, when comparing the models' knowledge coverage with that of radiologists, generalist models like RadFM and MedVersa show significantly broader node coverage. Note that here, all generalist models are trained on various tasks such as diagnosis, VQA, and report generation, but CheXagent only focuses on chest X-rays, while other generalist models include datasets from various modalities. From this, we can conclude that including data from various modalities improves the models' prediction generalizability, especially in terms of entity coverage. To develop medical AI systems that can interpret medical data and reason through complex problems at an expert radiologist level in real clinical scenarios, it is important to combine data from different modalities to broaden the models' knowledge base.

Ablation Studies

We conduct ablation studies to examine the impact of different medical embedding models, similarity thresholds, and the number of reports on our proposed metrics. The results are presented in Table 3. First, we compare the performance of two medical embedding models, BioLoRD (Remy, Demuynck, and Demeester 2024) and MedCPT (Jin et al. 2023), at different similarity thresholds. Our findings indicate that the choice of embedding model and threshold has a minimal effect on the extracted knowledge graph's quality. Both models perform robustly across different thresholds, with only slight variations in the KG-AMS metric. We also investigate how the number of reports influences the quality of the resulting knowledge graph. As expected, the number of reports significantly affects the results. However, we observe that as the number of reports increases, the performance asymptotically approaches that of the full dataset. For instance, with 10,000 studies, we achieve a KG-NSC of

Model	Threshold	# Study	KG-NSC	KG-AMS	KG-SCS
BioLoRD	0.95	24,085	**1.000**	**0.989**	**0.999**
BioLoRD	0.90	24,085	**1.000**	0.957	0.998
MedCPT	0.90	24,085	**1.000**	0.936	0.991
MedCPT	0.95	100	0.769	0.858	0.585
MedCPT	0.95	1,000	0.923	0.933	0.864
MedCPT	0.95	10,000	0.977	0.987	0.997

Table 3: Ablation studies on medical embedding models, similarity thresholds, and number of studies.

0.977 and a KG-AMS of 0.987, which closely matches the performance of the full dataset.

Related Work

Previous evaluations of radiology report generation models relied mainly on specific report-to-report metrics like FineRadScore (Huang et al. 2024), RaTEScore (Zhao et al. 2024), RadFact (Bannur et al. 2024), CheXPrompt (Chaves et al. 2024), and GREEN (Ostmeier et al. 2024). These metrics, however, do not fully capture an in-depth understanding of the capabilities of current models. Our work aims to address this limitation by leveraging knowledge graphs constructed from the report corpus. The standard pipeline for knowledge graph construction typically involves Named Entity Recognition (Li et al. 2020), Relation Extraction (Pawar, Palshikar, and Bhattacharyya 2017), and Entity Resolution (Christophides et al. 2020). In the medical domain, the focus has primarily been on developing knowledge graphs based on complex medical systems such as electronic health records, medical literature, and clinical guidelines (Rotmensch et al. 2017; Finlayson, LePendu, and Shah 2014; Bean et al. 2017). However, in the specific context of radiology reports, most progress focuses on information extraction (Irvin et al. 2019; McDermott et al. 2020; Peng et al. 2018; Smit et al. 2020; Jain et al. 2021b,a; Khanna et al. 2023; Delbrouck et al. 2024), and have not yet led to the establishment of a comprehensive knowledge graph specifically tailored for radiology reports. Few existing studies (Kale et al. 2022; Zhang et al. 2020) related to knowledge graph construction heavily relied on manual annotation by radiologists, highlighting the need for more automated, scalable approaches in this field.

Conclusion

In this paper, we present ReXKG, a novel system for constructing comprehensive radiology knowledge graphs from medical reports, and introduce three metrics for evaluating the similarity of nodes, distributions of edges, and coverage of subgraphs. We conduct an in-depth analysis comparing AI-generated radiology reports to human-written reports. Our research reveals that generalist models trained on various modalities offer broader coverage and enhanced radiology knowledge, yet they still fall short of the depth found in radiologist-written reports, particularly in the description and size measurements of disorders. Additionally, hallucinations related to prior studies are noticeable in model-

generated reports, highlighting the need to incorporate longitudinal data in future model development.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bannur, S.; Bouzid, K.; Castro, D. C.; Schwaighofer, A.; Bond-Taylor, S.; Ilse, M.; Pérez-García, F.; Salvatelli, V.; Sharma, H.; Meissen, F.; et al. 2024. MAIRA-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449*.

Bean, D. M.; Wu, H.; Iqbal, E.; Dzahini, O.; Ibrahim, Z. M.; Broadbent, M.; Stewart, R.; and Dobson, R. J. 2017. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports*, 7(1): 16416.

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

Chambon, P.; Delbrouck, J.-B.; Sounack, T.; Huang, S.-C.; Chen, Z.; Varma, M.; Truong, S. Q.; Chuong, C. T.; and Langlotz, C. P. 2024. CheXpert Plus: Hundreds of Thousands of Aligned Radiology Texts, Images and Patients. *arXiv preprint arXiv:2405.19538*.

Chaves, J. M. Z.; Huang, S.-C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; Awadalla, H. H.; Gong, J.; Hu, H.; Yang, J.; Li, C.; Gao, J.; Gu, Y.; Wong, C.; Wei, M.-H.; Naumann, T.; Chen, M.; Lungren, M. P.; Yeung-Levy, S.; Langlotz, C. P.; Wang, S.; and Poon, H. 2024. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Christophides, V.; Efthymiou, V.; Palpanas, T.; Papadakis, G.; and Stefanidis, K. 2020. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6): 1–42.

Delbrouck, J.-B.; Chambon, P.; Chen, Z.; Varma, M.; Johnston, A.; Blankemeier, L.; Van Veen, D.; Bui, T.; Truong, S.; and Langlotz, C. 2024. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In *Findings of the Association for Computational Linguistics ACL 2024*, 12902–12915.

Finlayson, S. G.; LePendu, P.; and Shah, N. H. 2014. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1(1): 1–9.

Huang, A.; Banerjee, O.; Wu, K.; Reis, E. P.; and Rajpurkar, P. 2024. FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. *arXiv preprint arXiv:2405.20613*.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.; Duong, D. N. D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.; Ng, A.; Langlotz, C.; Rajpurkar, P.; and Rajpurkar, P. 2021a. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jain, S.; Smit, A.; Truong, S. Q.; Nguyen, C. D.; Huynh, M.-T.; Jain, M.; Young, V. A.; Ng, A. Y.; Lungren, M. P.; and Rajpurkar, P. 2021b. VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, 105–115.

Jin, Q.; Kim, W.; Chen, Q.; Comeau, D. C.; Yeganova, L.; Wilbur, W. J.; and Lu, Z. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11): btad651.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kale, K.; Bhattacharyya, P.; Shetty, A.; Gune, M.; Shrivastava, K.; Lawyer, R.; and Biswas, S. 2022. Knowledge Graph Construction and Its Application in Automatic Radiology Report Generation from Radiologist's Dictation. *arXiv preprint arXiv:2206.06308*.

Khanna, S.; Dejl, A.; Yoon, K.; Truong, Q. H.; Duong, H.; Saenz, A.; and Rajpurkar, P. 2023. RadGraph2: Modeling Disease Progression in Radiology Reports via Hierarchical Information Extraction. *arXiv preprint arXiv:2308.05046*.

Li, J.; Sun, A.; Han, J.; and Li, C. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1): 50–70.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18635–18643.

Liu, C.; Tian, Y.; and Song, Y. 2023. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*.

McDermott, M. B.; Hsu, T. M. H.; Weng, W.-H.; Ghassemi, M.; and Szolovits, P. 2020. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, 913–927. PMLR.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Nicolson, A.; et al. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144: 102633.

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Michalson, A. E.; Moseley, M.; Langlotz, C.; Chaudhari, A. S.; et al. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. *arXiv preprint arXiv:2405.03595*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pawar, S.; Palshikar, G. K.; and Bhattacharyya, P. 2017. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.

Peng, Y.; Wang, X.; Lu, L.; Bagheri, M.; Summers, R.; and Lu, Z. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018: 188.

Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature medicine*, 28(1): 31–38.

Rajpurkar, P.; and Lungren, M. P. 2023. The current and future state of AI interpretation of medical images. *New England Journal of Medicine*, 388(21): 1981–1990.

Reale-Nosei, G.; et al. 2024. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, 103264.

Remy, F.; Demuynck, K.; and Demeester, T. 2024. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, ocae029.

Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; and Sontag, D. 2017. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1): 5994.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC2 Workshop*.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167*.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. *arXiv preprint arXiv:2308.02463*.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021a. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22–31.

Wu, J. T.; Agu, N. N.; Lourentzou, I.; Sharma, A.; Paguio, J. A.; Yao, J. S.; Dee, E. C.; Mitchell, W.; Kashyap, S.; Giovannini, A.; et al. 2021b. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Lee, H. M. H.; Abad, Z. S. H.; Ng, A. Y.; et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; and Xu, D. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12910–12917.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A Metric for Radiology Report Generation. *medRxiv*, 2024–06.

Zhong, Z.; and Chen, D. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 50–61.

Zhou, H.-Y.; Adithan, S.; Acosta, J. N.; Topol, E. J.; and Rajpurkar, P. 2024. A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

Appendix

Prompt for Entity Extraction

You are a radiologist performing clinical term extraction from the FINDINGS and IMPRESSION sections in the radiology report. Here a clinical term can be in [`anatomy`, `disorder_present`, `disorder_notpresent`, `procedure`, `device_present`, `device_notpresent`, `size`, `concept`]. `anatomy` refers to the anatomical body. `disorder_present` refers to findings or diseases that are present according to the sentence. `disorder_notpresent` refers to findings or diseases that are not present according to the sentence. `procedure` refers to procedures used to diagnose, measure, monitor, or treat problems. `device_present` refers to any instrument, apparatus for medical purpose that are present according to the sentence. `device_notpresent` refers to any instrument, apparatus for medical purpose that are not present according to the sentence. `size` refers to the measurement of disorders or anatomy, for example, `3mm`, `4x5 cm`. `concept` refers to descriptors such as `acute` or `chronic`, `large`, size or severity, or other modifiers, or descriptors of anatomy being normal. For example, right pleural effusion, `right` should be a `concept`, and `pleural` should be `anatomy` and `effusion` should be `disorder-present` or `disorder-notpresent`. For example, normal cardiomediastinal silhouette. `normal` and `silhouette` should be `concept`, `cardiomediastinal` should be `anatomy`. Please extract terms one word at a time whenever possible, avoiding phrases. Note that terms like `no` and `no evidence of` are not considered entities. Given a list of radiology sentence input in the format: <Input><sentence><sentence></Input> Please reply with the JSON format following template: {<sentence>{entity:entity type, entity:entity type}, <sentence>{entity:entity type, entity:entity type}}.

Prompt for Relation Extraction

You are a radiologist performing relation extraction of entities from the FINDINGS and IMPRESSION sections in the radiology report. Here a clinical term can be in [`anatomy`, `disorder_present`, `disorder_notpresent`, `procedures`, `procedures`, `concept`, `devices_present`, `devices_notpresent`]. And the relation can be in [`modify`, `located_at`, `suggestive_of`]. `suggestive_of` means the source entity (findings) may suggest the target entity (disease). `located_at` means the source entity is located at the target entity. `modify` denotes the source entity modifies the target entity. Every time there is a `modify` relationship

between concept and anatomy, the direction should be concept → anatomy. For example, paranasal sinuses are clear: source entity `clear` (concept), modify target entity `paranasal sinuses` (anatomy). For example, acute hemorrhage: source entity `acute` (concept), modify target entity `hemorrhage`. Given a piece of radiology text input in the JSON format: {sentence:{entity:entity_type}, sentence:{entity:entity_type}}. Please reply with the following JSON format: {sentence:[{source entity:target entity, relation:relation}, {source entity:target entity, relation:relation}}

Algorithm for Node Construction

Algorithm 1: Node Integration

Require: E: list of entities
Require: C: count threshold
Require: n: maximum number of words in an entity
1: Initialize $A \leftarrow \emptyset$ {Set of initial nodes}
2: Group E by word count and filter by C
3: **for** each k from 1 to n **do**
4:   **for** each $e \in E$ with k words **do**
5:     **if** $k == 1$ **then**
6:       Add e to set A
7:     **else**
8:       **if** e can merge from nodes in A **then**
9:         Pass
10:       **else**
11:         Add e to set A
12:       **end if**
13:     **end if**
14:   **end for**
15: **end for**
16: **return** A {Set of nodes}

KG Subgraph Coverage Score

Let $\mathcal{S} = \{S_1, S_2, ..., S_L\}$ be the set of all connected subgraphs in KG-GT with the size of k nodes. For each subgraph S_i, we compute an importance score $I(S_i)$ based on the frequency of occurrence and total edge weights:

$$I(S_i) = \sum_{v \in V(S_i)} w_v \cdot \sum_{e \in E(S_i)} w_e, \qquad (4)$$

where $V(S_i)$ and $E(S_i)$ denote the vertex and edge sets of S_i respectively, and w_v and w_e are the corresponding node and edge weights. For each subgraph S_i in KG-GT, we compute a presence score $P(S_i)$ in KG-Pred:

$$P(S_i) = \frac{1}{2} \left(\frac{|E(S_i')|}{|E(S_i)|} + \frac{\sum_{v \in V(S_i)} s_v}{|V(S_i)|} \right), \qquad (5)$$

where S_i' is the corresponding subgraph in KG-Pred, $|E(.)|$ and $|V(.)|$ denote the number of edges and vertices respectively, and s_v is the similarity score between matched nodes

Dataset	Source	Target	KG-NSC						KG-AMS				KG-SCS
			Ana.	Dis.	Con.	Dev.	Pro.	All	Dis.Ana.	Dev.Ana.	Dis.Dis.	All	k=2
CT-RATE	Part I	Part II	0.977	0.971	0.984	0.955	0.973	0.978	0.997	0.914	0.972	0.974	0.999
CT-RATE	Part II	Part I	0.982	0.968	0.977	0.977	0.991	0.977	0.997	0.974	0.993	0.948	0.998
MIMIC-IV Head CT	Part I	Part II	0.986	0.976	0.986	0.976	0.987	0.984	0.989	0.986	0.994	0.993	0.999
MIMIC-IV Head CT	Part II	Part I	0.981	0.977	0.983	0.952	0.972	0.980	0.994	0.987	0.996	0.987	0.999

Table A1: Knowledge graph comparison on CT-RATE and MIMIC-IC Head CT datasets. KG-NSC, KG-AMS, and KG-SCS scores are reported. The best results are highlighted in boldface.

Type	Models	KG-NSC						KG-AMS				KG-SCS
		Ana.	Dis.	Con.	Dev.	Pro.	All	Dis.Ana.	Dev.Ana.	Dis.Dis.	All	k=2
Intra-Dataset	CheXpert Plus I	0.970	0.967	0.974	0.980	0.980	0.973	0.954	0.983	0.985	0.968	0.997
	MIMIC-CXR	0.920	0.956	0.936	0.882	0.938	0.932	0.844	0.807	0.849	0.832	0.952
Specialist	CvT2DistilGPT2 (Nicolson et al. 2023)	0.776	0.772	0.787	0.747	0.806	0.781	0.751	0.846	0.692	0.644	0.664
	RGRG (Tanida et al. 2023)	0.664	0.636	0.618	0.597	0.568	0.626	0.612	0.681	0.725	0.578	0.529
	Swinv2-MIMIC (Chambon et al. 2024)	0.790	0.792	0.774	0.732	0.812	0.780	0.690	0.811	0.719	0.660	0.625
Generalist	CheXagent (Chen et al. 2024)	0.715	0.696	0.698	0.686	0.718	0.702	0.779	**0.877**	0.566	0.711	0.555
	RadFM (Wu et al. 2023)	**0.804**	**0.831**	0.788	0.728	0.765	0.792	0.681	0.704	0.613	0.615	0.635
	MedVersa (Zhou et al. 2024)	**0.804**	0.824	**0.800**	**0.750**	**0.813**	**0.802**	**0.800**	0.851	**0.893**	**0.723**	**0.709**

Table A2: Knowledge graph comparison between CheXpert Plus II and Intra-Dataset or Extra-Dataset Reports. KG-NSC, KG-AMS, and KG-SCS scores are reported. The best results are highlighted in boldface.

as defined in the KG-NSC section. The Subgraph Coverage Score is then calculated as:

$$\text{KG-SCS} = \frac{\sum_{i=1}^{K} I(S_i) \cdot P(S_i)}{\sum_{i=1}^{K} I(S_i)}, \qquad (6)$$

where K is the number of top important subgraphs considered, $I(S_i)$ is the normalized importance score of subgraph S_i among the selected K subgraphs.

Report Genertaion Models

- **CvT2DistilGPT2** (Nicolson et al. 2023): The model adopts the Convolutional Vision Transformer (CvT) (Wu et al. 2021a) pre-trained on ImageNet-21K (Ridnik et al. 2021) for the visual encoder and the Distilled Generative Pre-trained Transformer 2 (DistilGPT2) (Sanh et al. 2019) for the text decoder. We use the released checkpoint trained on the MIMIC-CXR dataset.

- **RGRG**(Tanida et al. 2023): The model employs an anatomy-based object detector, fine-tuned on the Chest ImaGenome dataset (Wu et al. 2021b), which identifies 29 annotated anatomical regions. These regional visual features are then used to guide the generation of detailed and clinically relevant radiology reports

- **Swinv2-MIMIC**(Chambon et al. 2024): The model is proposed as a baseline model for report generation on the CheXpert Plus dataset (Chambon et al. 2024). It builds upon the Swin Transformer architecture, and for our experiments, we use the released checkpoint trained on the MIMIC-CXR findings dataset.

- **CheXagent**(Chen et al. 2024): The model is trained on the CheXinstruct dataset, which utilizes a clinical large language model for parsing radiology reports, a vision encoder for CXR representation, and a network that bridges vision and language modalities.

- **RadFM**(Wu et al. 2023): The model is a radiology foundation model trained on large-scale multi-modal medical datasets, which enables the integration of text input interleaved with 2D or 3D medical scans to generate responses for diverse radiologic tasks.

- **MedVersa**(Zhou et al. 2024): The model is a versatile model trained on large-scale medical data across multiple modalities and tasks, which supports multimodal inputs, outputs, and on-the-fly task specification.

Evaluation Metrics

- **BLEU** (Papineni et al. 2002) evaluates the precision of generated text by comparing n-gram overlap between the generated report and reference reports.

- **BERTScore** (Zhang et al. 2019) employs a pre-trained BERT model to compute the similarity of word embeddings between candidate and reference texts.

- **SembScore** (Smit et al. 2020) refers to the CheXbert labeler vector similarity. This method uses a 14-dimensional vector to indicate the presence of 13 common symptoms and the "no finding" observation for each report, then calculates the cosine similarity between these vectors.

- **RadGraph F1** (Jain et al. 2021a) extracts radiology entities and relations specifically for Chest X-ray modality and computes the F1 score at the entity level.
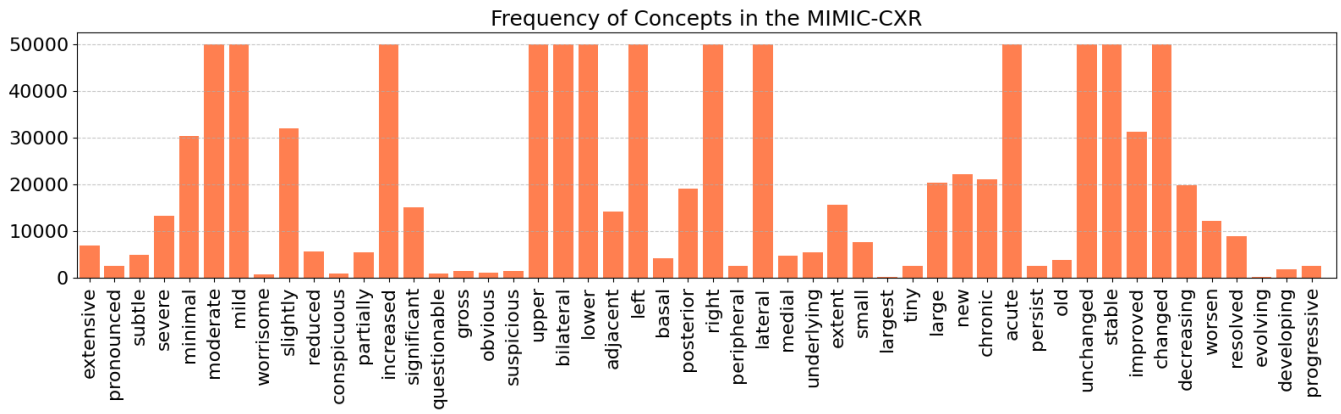
Figure A1: Frequency of concepts used to modify different disorders in the training set MIMIC-CXR.

- **RadCliQ-v1** (Yu et al. 2023) is a composite metric that incorporates BLEU, BERTScore, SembScore, and RadGraph F1.

Demonstration of ReXKG on various modalities

The proposed knowledge graph construction system is versatile and can be applied across various modalities and anatomical regions. We further demonstrate its effectiveness on CT-RATE and MIMIC-IV Head CT reports, similar to the chest x-ray experiments. For these studies, we randomly split the target dataset into two equal parts and compared the knowledge graphs constructed from each subset.

- **CT-RATE**: CT-RATE consists of 25,692 non-contrast chest CT volumes, expanded to 50,188 through various reconstructions, from 21,304 unique patients, along with corresponding radiology text reports. Here, we split the studies into two parts, Part I and Part II.
- **MIMIC-IV Head CT**: MIMIC-IV notes include reports from various modalities. Here we select the reports from head CT, including 101,633 studies, and split them into two parts, Part I and Part II.

The results in Table A1 show that when two corpora used for knowledge graph construction are of similar quality, the scores are consistently high. This indicates that the metrics are robust and suitable for evaluating knowledge graphs across various modalities.

Results with CheXpert Plus II as benchmark

Here, we set CheXpert Plus II as the benchmark and reproduce all the experiments, with results provided in Table A2. As shown, the experimental results are consistent with those presented in the results section using CheXpert Plus I as the benchmark.

Analysis of the concept used to modify disorders

Figure A1 illustrates the frequency distribution of the analyzed concepts in the MIMIC-CXR training set. Figure A2 depicts the frequency of size descriptions for specific disorders in the MIMIC-CXR training data. Figure A3 and Figure A4 provide comprehensive results on high-frequency
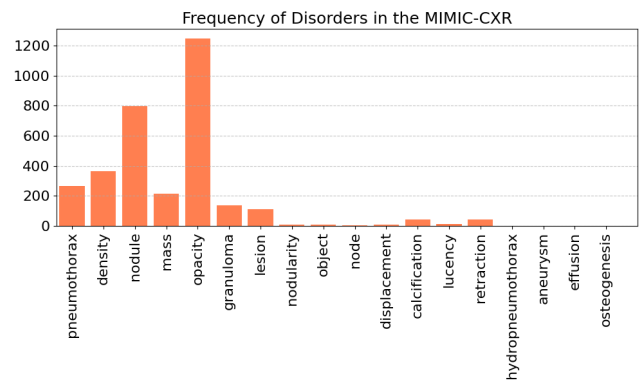

Figure A2: Frequency of size measurement for different disorders in the training set MIMIC-CXR.

disorders and the commonly used concepts to modify these disorders across different models.
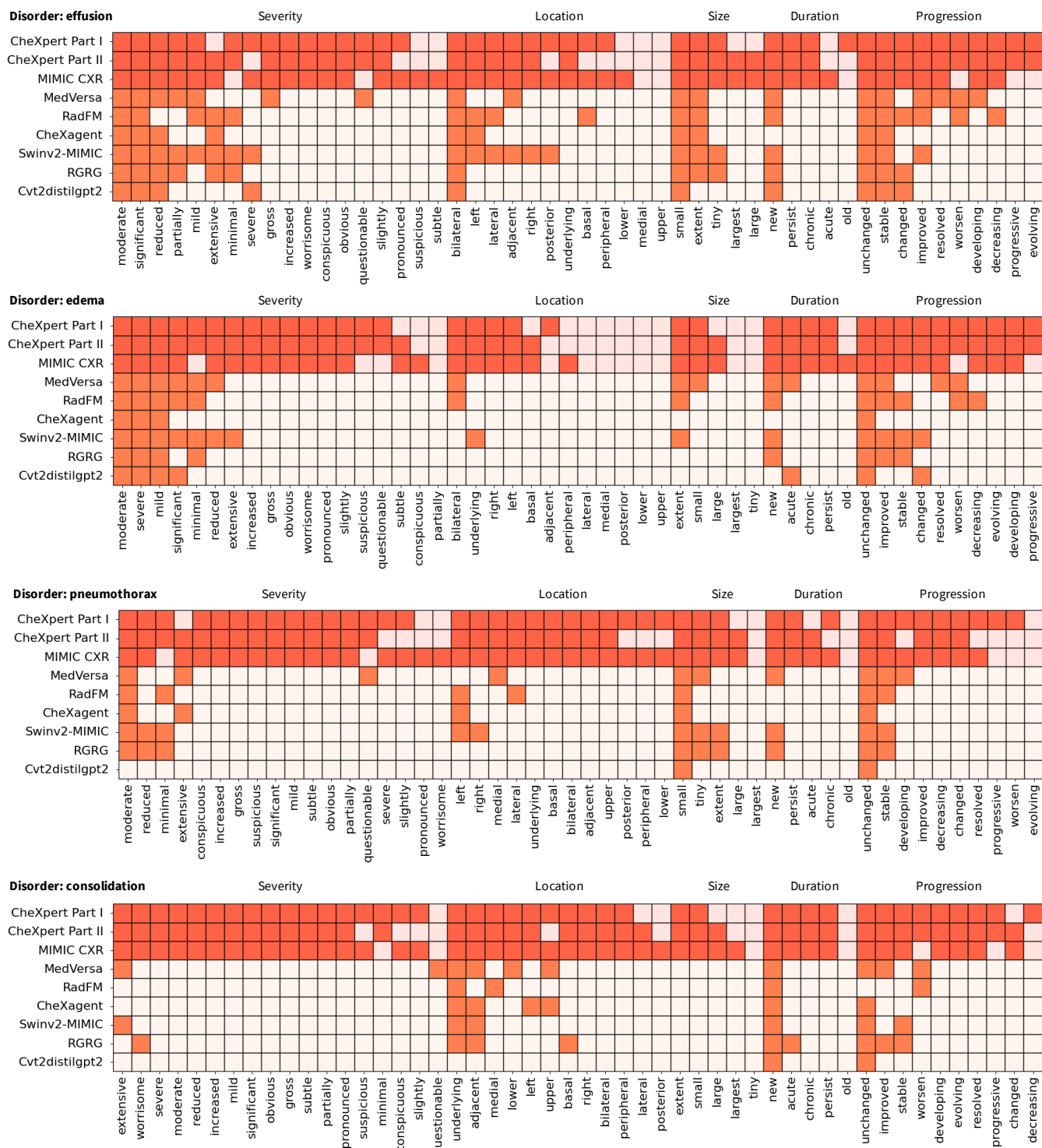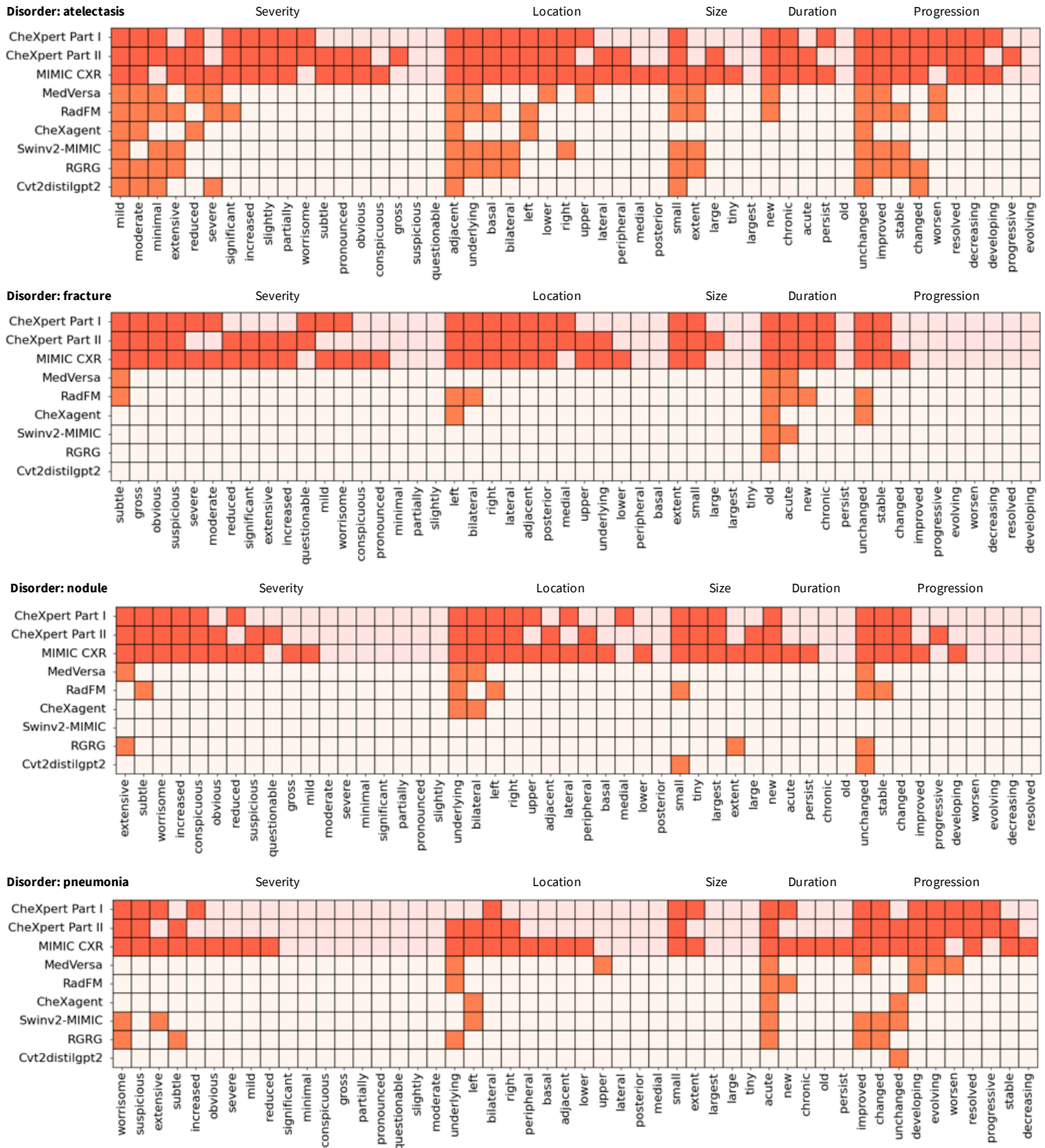
Figure A3: Detailed results of model predictions.

Figure A4: Detailed results of model predictions.