Uncovering Knowledge Gaps in Radiology Report Generation Models through Knowledge Graphs

Xiaoman Zhang XIAOMAN_ZHANG@HMS.HARVARD.EDU
Julián N. Acosta JULIAN_ACOSTA@HMS.HARVARD.EDU
Hong-Yu Zhou HONGYU_ZHOU@HMS.HARVARD.EDU
Pranav Rajpurkar * PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Abstract

Recent advancements in artificial intelligence have significantly improved the automatic generation of radiology reports. However, existing evaluation methods often focus on report-to-report similarities and 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, distribution of edges, and coverage of subgraphs across various knowledge graphs. Using these metrics, 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 report generation, offering valuable insights for improving model performance and clinical applicability.

Data and Code Availability This paper uses the CheXpert Plus (Chambon et al., 2024) and MIMIC-CXR (Johnson et al., 2019). The MIMIC-CXR dataset is available on the PhysioNet repository (https://physionet.org/content/mimic-cxr-jpg/2.1.0/). The CheXpert Plus dataset is available at (https://aimi.stanford.edu/datasets/chexpert-plus). The official code is available at https://github.com/rajpurkarlab/ReXKG.

Institutional Review Board (IRB) This research uses publicly available data sets (MIMIC-CXR and CheXpert Plus) and does not involve direct research of human subjects. Therefore, IRB approval was not required for this study.

1. 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 et al., 2023a; 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

* Corresponding Author

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 concepts 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.

2. 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 et al., 2017), and Entity Resolution (Christophides et al., 2020). With the remarkable capabilities in natural language processing, Large Language Models are increasingly utilized for creating and enriching comprehensive ontologies, facilitating information extraction, and enabling information prediction (Han et al., 2023; Neuhaus, 2023; Ashok and Lipton, 2023; Khorashadizadeh et al., 2024). 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 et al., 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

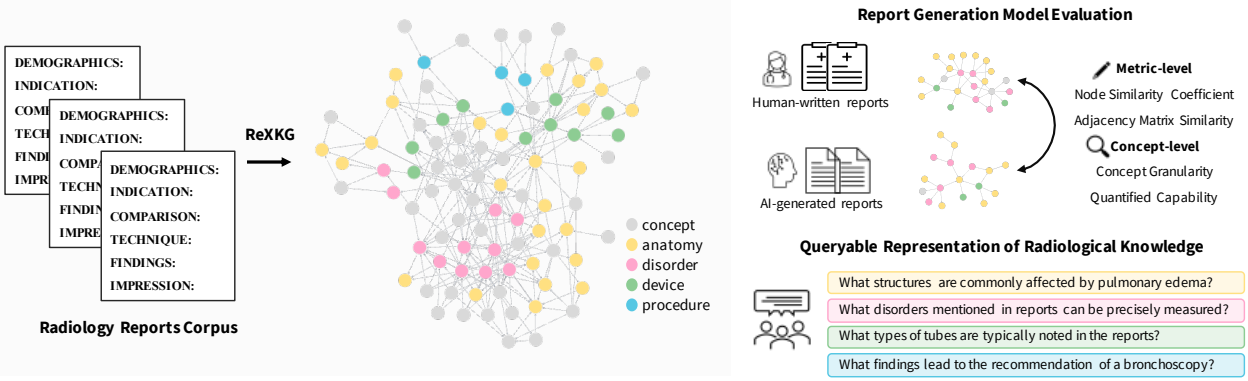

Figure 1: An illustration of **Learning from Knowledge Graph**. We propose ReXKG to construct a knowledge graph that includes concept, anatomy, disorder, device, and procedure nodes. Based on this, we can evaluate report generation models at multiple levels, and enable queryable representations of radiological knowledge, facilitating detailed radiology-related inquiries.

the need for more automated, scalable approaches in this field.

3. 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.

3.1. Information Extraction Schema

Definition. While previous works such as RadGraph (Jain et al., 2021a) have proposed schemas for radiology report annotation, they categorize entities into broad types like "Observation" and "Anatomy" but do not distinguish between devices, disorders, and concepts. This lack of granularity is insufficient for capturing the full range and depth of information needed for comprehensive analysis. We define an entity as a continuous span of text that can include one or more adjacent words. Our schema categorizes entities into six types: **Anatomy**, **Disorder**, **Concept**, **Device**, **Procedure**, and **Size**. A relation is defined as a directed edge between two entities, utilizing three types: **Suggestive of**, **Located at**, and **Modify**. Detailed definitions and examples for each category are provided in the appendix.

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.

3.2. 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 assign each entity its most frequent type as predicted by the NER model across all occurrences in the dataset, ensuring type consistency. 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

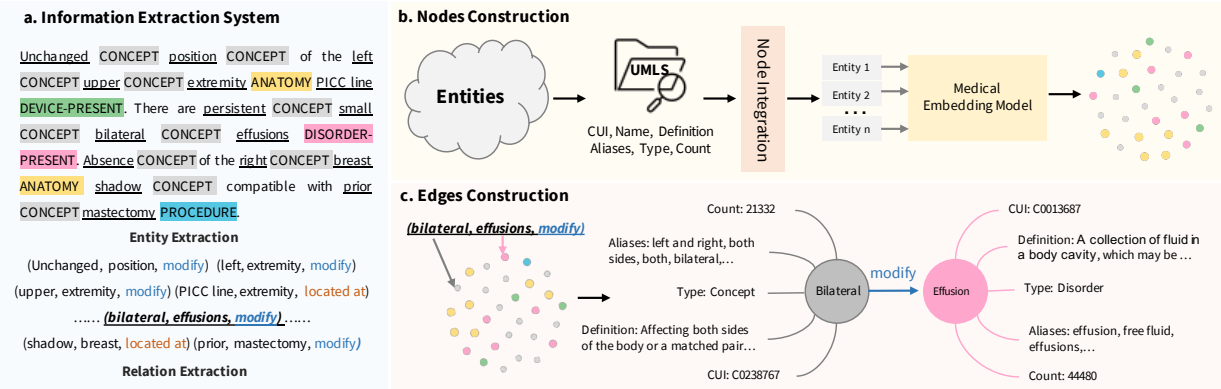

Figure 2: Overview of the proposed knowledge graph construction system **ReXKG**. (a) The information extraction system for entity and relation extraction. (b) The node construction pipeline. (c) Illustration of edge construction.

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 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.

3.3. 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. This process removes only the edges meeting this criterion while retaining the nodes in the graph.

4. 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 proposed metrics.

4.1. 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 $w_{e_m} = \text{count}(e_m)$. 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.

4.2. KG Node Similarity Coefficient

Let KG-GT represent the knowledge graph built from the ground truth reports, consisting of N nodes. Sim-

ilarly, 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 embeddings generated by a medical language model (MedCPT (Jin et al., 2023), by default). 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:

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

4.3. 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:

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

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.

4.4. 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:

$$I(S_i) = \sum_{v \in V(S_i)} w_v \cdot \sum_{e \in E(S_i)} w_e, \tag{3}$$

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:

$$P(S_i) = \frac{1}{2}\left(\frac{|E(S_i')|}{|E(S_i)|} + \frac{\sum_{v \in V(S_i)} s_v}{|V(S_i)|}\right), \tag{4}$$

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

$$\texttt{KG-SCS} = \frac{\sum_{i=1}^{K} I(S_i) \cdot P(S_i)}{\sum_{i=1}^{K} I(S_i)}, \tag{5}$$

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.

5. 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.

5.1. 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 (License: Stanford University Dataset Research Use Agreement). 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 (License: PhysioNet Credentialed Health Data License 1.5.0). The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center.

5.2. 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.

5.3. 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 pre-trained 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 MedCPT (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.

6. 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 multifaceted view of current AI models' capabilities.

6.1. 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 1, 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 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

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. Ana. refers to Anatomy, Dis. refers to disorder, Con. refers to Concept, Dev. refers to Device, Pro. refers to Procedure. The best results are highlighted in boldface.

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".

6.2. Coverage of Relationships Between Entities

Next, we investigate **To what extent do the AI-generated reports accurately represent specific clinical 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 1 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.

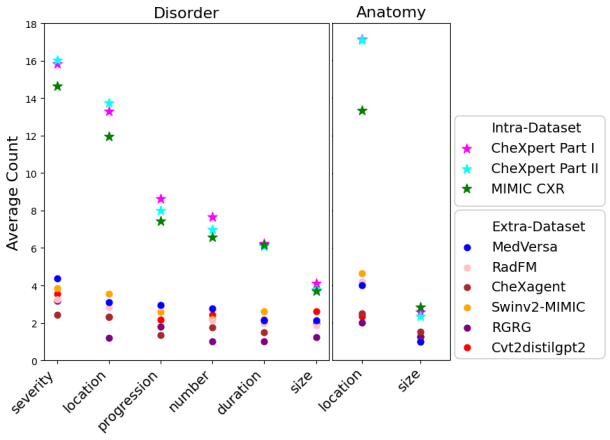

Figure 3: Average count of concept entities used to modify disorders and anatomy across different models.

6.3. Comprehensiveness of Concepts

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

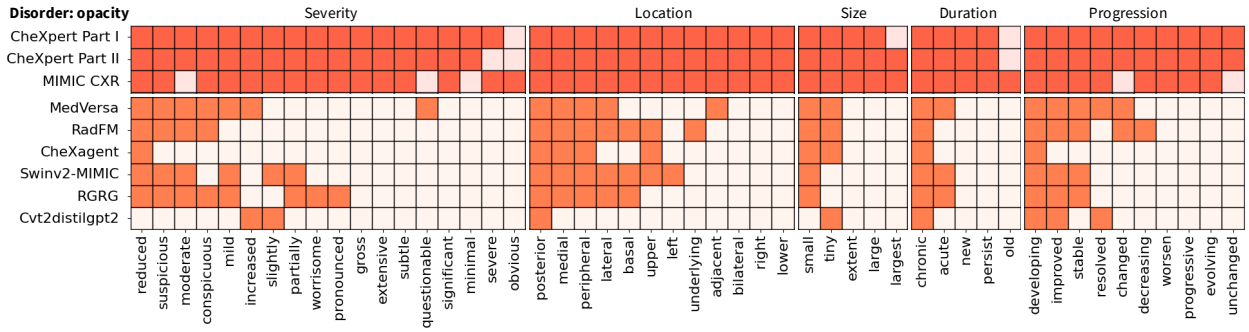

Figure 4: Detailed results of model predictions for given concepts related to specific disorders. Dark orange indicates the model predicts the relationship at least once in the dataset, while light orange indicates never.

pact 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 severity, location, duration, progression, size, and number (detailed definition of each category can be found in the appendix). 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. In addition, during our evaluation process, we observed that AI-generated reports demonstrate higher internal consistency compared to human-written reports. AI models tend to generate more structurally similar outputs, with less variation in descriptors, while different radiologists often employ distinct writing styles.

6.4. Quantitative Measurements Coverage

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

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

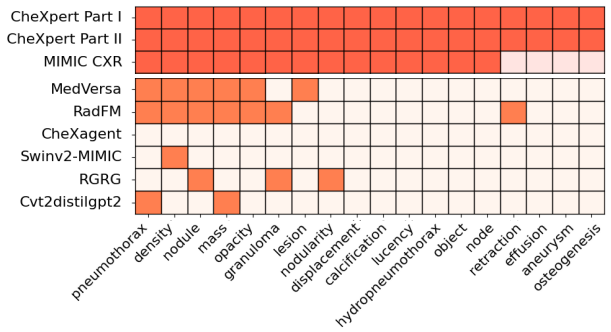

Figure 5: Detailed results of whether the model predicts specific size measurements for given disorders. Dark orange indicates the model predicts size measurement for the target disorder at least once in the dataset, while light orange indicates never.

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.

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.

6.5. Specialist vs. Generalist Models

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

We summarize the score of different metrics on different models' predictions on the training set of CheXpert Plus finding sections in Table 2. 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.

6.6. 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 et al., 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

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.

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

7. Conclusion

In this paper, we present ReXKG, a novel system for constructing radiology knowledge graphs from medical reports, and introduce three metrics for evaluating the node similarity, edges distributions, and subgraph coverage. These enable us to 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.

8. Limitations

The core contribution of this paper is the use of knowledge graphs to compare reports generated by different report-generation models. While this approach offers novel insights, it still exhibits several limitations. (1) The direct evaluation of the constructed knowledge graphs is challenging. We attempt to demonstrate the effectiveness of our Knowledge Graph Construction method by comparing metric scores between Part I and Part II of the CheXpert Plus dataset. Our underlying assumption is that two datasets with similar distributions should produce high similarity scores. However, this indirect validation approach may not comprehensively capture all aspects of graph quality and could potentially overlook nuanced differences that are clinically significant. (2) The performance of the NER models significantly influences the final accuracy of the knowledge graphs. While large language models have shown promising results in information extraction (Liu et al., 2023b), we did not utilize them in the final pipeline due to computational cost constraints. Future work could explore integrating these models as they become more cost-effective, potentially enhancing the accuracy of our knowledge graph construction process. (3) Although our proposed method is theoretically applicable to all modalities, our experiments and analysis primarily focus on chest X-ray reports. This narrow scope, partly due to current model limitations, may restrict the generalizability of our findings to other radiological modalities or anatomical regions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Dhananjay Ashok and Zachary C Lipton. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*, 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Daniel M Bean, Honghan Wu, Ehtesham Iqbal, Olubanke Dzahini, Zina M Ibrahim, Matthew Broadbent, Robert Stewart, and Richard JB Dobson. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports*, 7(1):16416, 2017.

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

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and

Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. 2024. URL https://api.semanticscholar.org/CorpusID:268379244.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6):1–42, 2020.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. 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*, pages 12902–12915, 2024.

Samuel G Finlayson, Paea LePendu, and Nigam H Shah. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1(1): 1–9, 2014.

Jiuzhou Han, Nigel Collier, Wray Buntine, and Ehsan Shareghi. Pive: Prompting with iterative verification improving graph-based generative capability of llms. *arXiv preprint arXiv:2305.12392*, 2023.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv preprint arXiv:2405.20613*, 2024.

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*, volume 33, pages 590–597, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021a. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021b.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer, and Spriha Biswas. Knowledge graph construction and its application in automatic radiology re-

port generation from radiologist's dictation. *arXiv preprint arXiv:2206.06308*, 2022.

Sameer Khanna, Adam Dejl, Kibo Yoon, Quoc Hung Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. *arXiv preprint arXiv:2308.05046*, 2023.

Hanieh Khorashadizadeh, Fatima Zahra Amara, Morteza Ezzabady, Frédéric Ieng, Sanju Tiwari, Nandana Mihindukulasooriya, Jinghua Groppe, Soror Sahri, Farah Benamara, and Sven Groppe. Research trends for the interplay between large language models and knowledge graphs. *arXiv preprint arXiv:2406.08223*, 2024.

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

Chang Liu, Yuanhe Tian, and Yan Song. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*, 2023a.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643, 2024.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. Exploring the boundaries of gpt-4 in radiology. *arXiv preprint arXiv:2310.14573*, 2023b.

Matthew BA McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR, 2020.

Fabian Neuhaus. Ontologies in the era of large language models–a perspective. *Applied ontology*, 18 (4):399–407, 2023.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

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

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*, 2017.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

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

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

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

François Remy, Kris Demuynck, and Thomas Demeester. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, page ocae029, 2024.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):5994, 2017.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiol-

ogy report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

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*, volume 34, pages 12910–12917, 2020.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *medRxiv*, pages 2024–06, 2024.

Zexuan Zhong and Danqi Chen. 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*, pages 50–61, 2021.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.