Appendix A. Appendix

A.1. 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`}}.

A.2. 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`}}

A.3. Detailed Definition of Information Extraction Schema

Entities in our schema are categorized into six types as listed.

- **Anatomy**: anatomical structures.
- **Disorder**: any abnormal findings or diseases identified within radiology reports.
- **Concept**: descriptors used to modify other entities, for example, “acute”, “severe”
- **Device**: any instrument or apparatus used for medical purposes, for example, “tube”, “clip”.
- **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".

Relations are categorized 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.

A.4. 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
Initialize $A \leftarrow \emptyset$ {Set of initial nodes}
Group E by word count and filter by C
for each k from 1 to n **do**
  for each $e \in E$ with k words **do**
    if $k == 1$ **then**
      Add e to set A
    else
      if e can merge from nodes in A **then**
        Pass
      else
        Add e to set A
      end if
    end if
  end for
end for
return A {Set of nodes}

A.5. 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.

A.6. 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.

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.

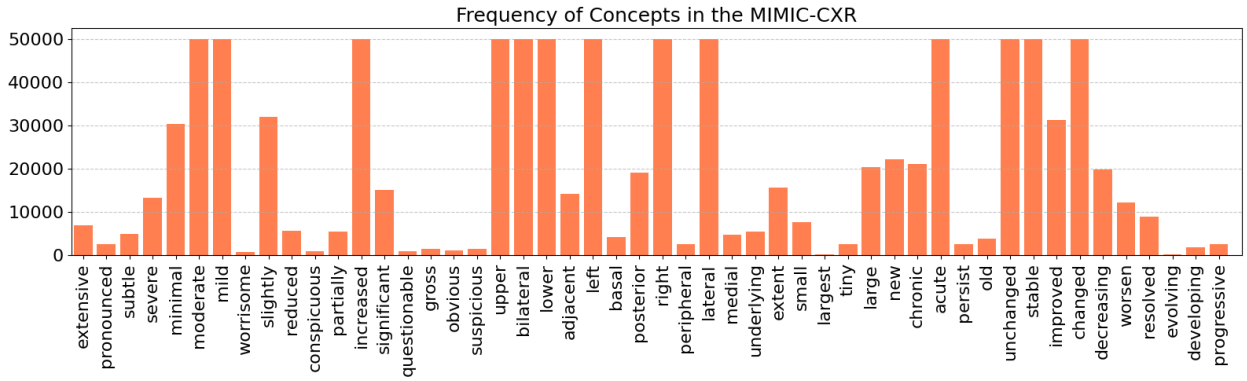

Figure A1: Frequency of concepts used to modify different disorders in the training set MIMIC-CXR.

- **RadGraph F1** (Jain et al., 2021) extracts radiology entities and relations specifically for Chest X-ray modality and computes the F1 score at the entity level.

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

A.7. Concept Categories

- **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.

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

A.8. 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 (License: Creative Commons Attribution Non-Commercial Share Alike 4.0). 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 (License: PhysioNet Credentialed Health Data License 1.5.0). 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.

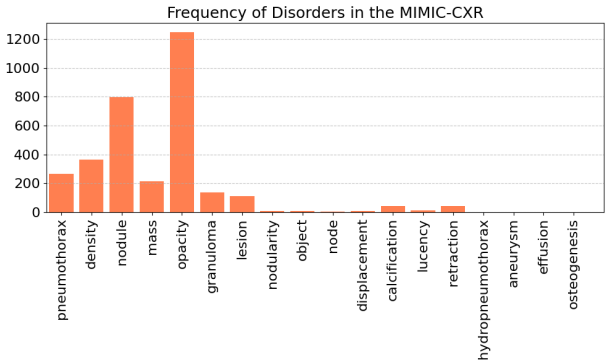

Figure A2: Frequency of size measurement for different disorders in the training set MIMIC-CXR.

A.9. 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.

A.10. 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 disorders and the commonly used concepts to modify these disorders across different models.

References

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.

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.

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, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC2 Workshop*, 2019.

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

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021a.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021b.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

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.

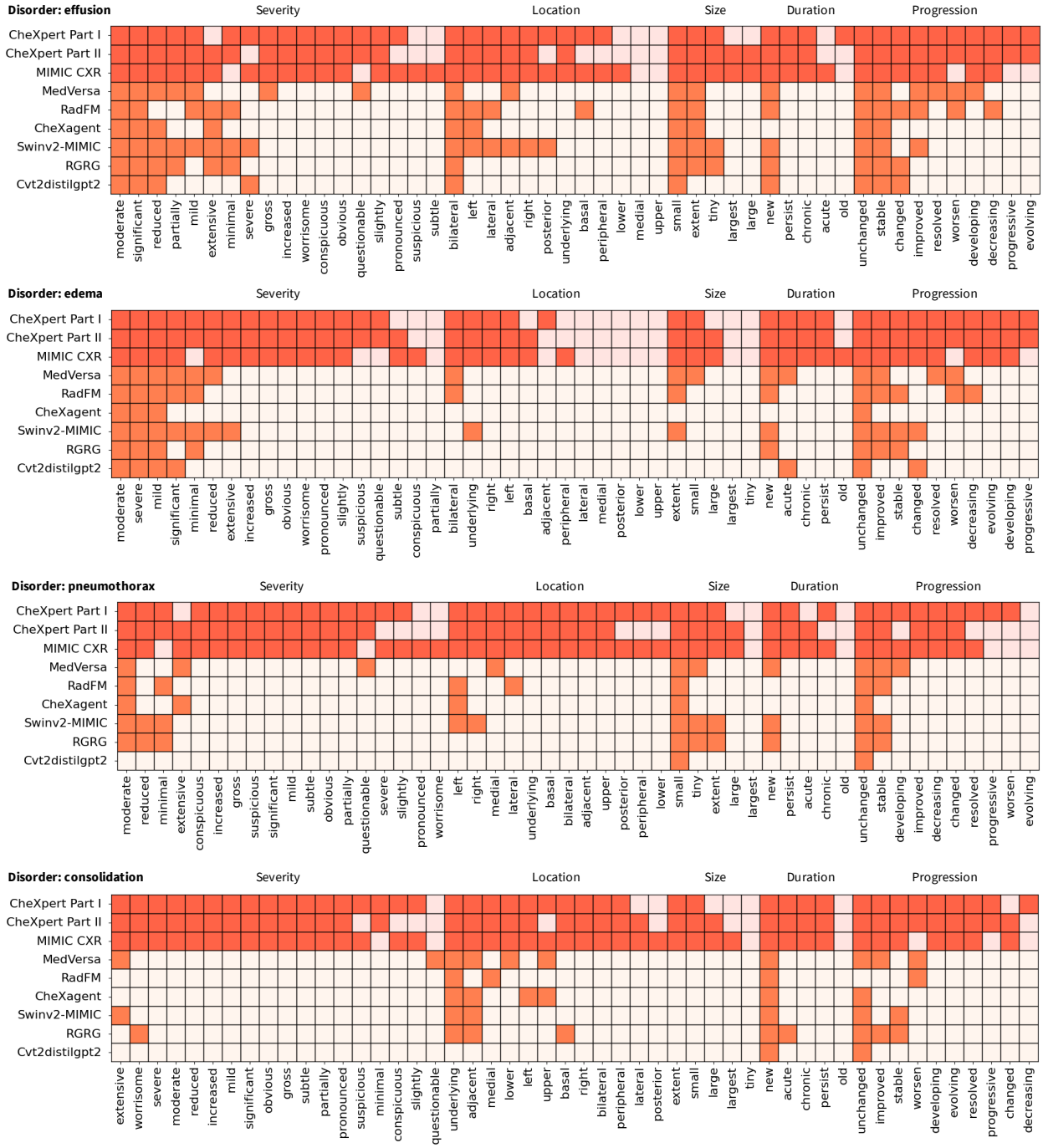

Figure A3: Detailed results of model predictions.

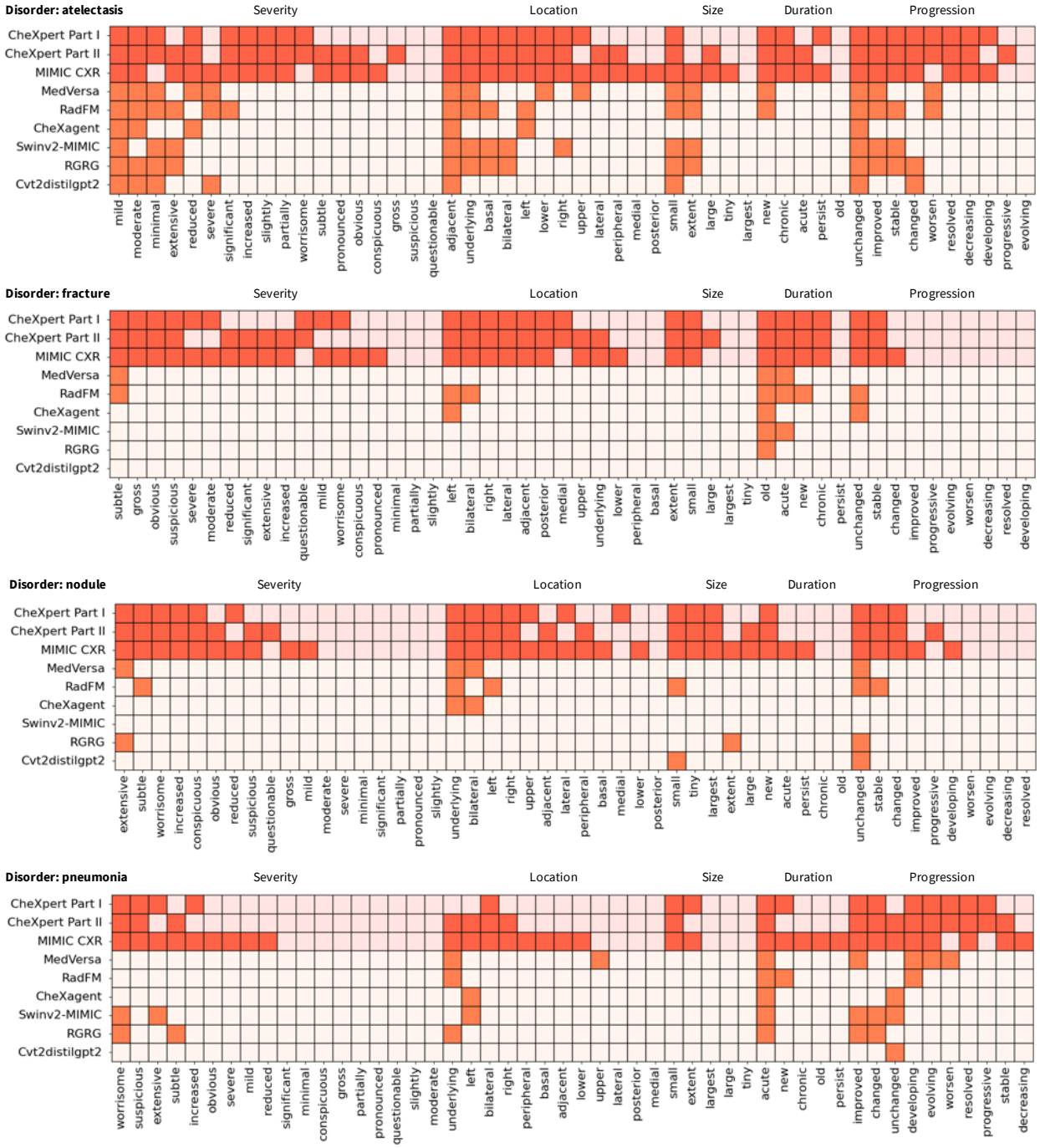

Figure A4: Detailed results of model predictions.