Supplementary Material

A Measures

For both experiments and model variants, we formulate each clinical diagnosis task as a multi-label classification problem over all possible labels in a dataset. The evaluation metrics are computed as follows:

$$\text{Accuracy}_{\text{label}} = \frac{TP+TN}{TP+TN+FP+FN}, \qquad \text{Precision}_{\text{label}} = \frac{TP}{TP+FP}, \qquad \text{Recall}_{\text{label}} = \frac{TP}{TP+FN} \tag{1}$$

$$\text{F1}_{\text{label}} = 2 \cdot \frac{\text{Precision}_{\text{label}} \cdot \text{Recall}_{\text{label}}}{\text{Precision}_{\text{label}} + \text{Recall}_{\text{label}}} \tag{2}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively, for each label across all samples. To account for potential class imbalance, we compute the balanced accuracy for each dataset:

$$\text{Balanced Accuracy}_{\text{dataset}} = \frac{1}{|L|} \sum_{l \in L} \text{Accuracy}_{\text{label}} \tag{3}$$

where L is the set of all labels in the dataset. Similarly, we compute the macro-F1 score for each dataset:

$$\text{Macro-F1}_{\text{dataset}} = \frac{1}{|L|} \sum_{l \in L} \text{F1}_{\text{label}} \tag{4}$$

For open-ended questions, we use Jaccard similarity to measure the lexical overlap between the predicted response and the ground-truth answer. Let $W_{\hat{y}}$ denote the set of unique words in the predicted response $\hat{y}$ and W_y denote the set of unique words in the ground-truth answer y, both after lowercasing and removing stopwords. The Jaccard similarity is defined as:

$$\text{Jaccard Similarity} = \frac{|W_{\hat{y}} \cap W_y|}{|W_{\hat{y}} \cup W_y|} \tag{5}$$

This metric provides a continuous score in $[0, 1]$ proportional to the lexical overlap between the prediction and ground truth, making it suitable for evaluating free-form textual responses.

To obtain domain-level metrics, we perform unweighted averaging across all datasets within each clinical domain:

$$\text{Metric}_{\text{domain}} = \frac{1}{|D_{\text{domain}}|} \sum_{d \in D_{\text{domain}}} \text{Metric}_{\text{dataset}} \tag{6}$$

where D_{domain} represents the set of datasets in a specific clinical domain, and Metric can be either balanced accuracy, macro-F1, or Jaccard similarity depending on the question type. Finally, we compute the overall performance by averaging the domain-level metrics across all 9 clinical domains:

$$\text{Metric}_{\text{overall}} = \frac{1}{9} \sum_{i=1}^{9} \text{Metric}_{\text{domain}_i} \tag{7}$$

B Training Hyperparameters.

Unless mentioned otherwise, we use the same set of hyperparameters to train the model across different training methods. The models are trained for 1 epoch on an 8xNVIDIA A100 and H200 GPU instances. For 7B models, we use a per-device batch size of 4, and a rollout batch size of 512. The maximum context length is 8192. To ensure consistency throughout the training, we shuffle the data with seed 42 beforehand, and disable shuffling throughout the training process. To save compute, we employ early stopping, which stops training when the accuracy converges and stops improving. Most 7B model trainings converge within 2 days of training. The training of 32B model takes more than 2 weeks to train on an 8xA100 machine, so a 8xH200 machine is used to speed up the training process of 32B model via faster interconnect.

We build our training pipeline based on the FSDP and VeRL framework, with vLLM to speed up reasoning training with KV Cache. We use a learning rate of 1e-6, a weight decay of 1e-2, and a KL coefficient of 1e-4. We use AdamW full model training at 32-bit precision for all 7B models, and at 16-bit precision for the training of the 32B model.

Throughout the training and evaluation, the 3D images and videos are sliced uniformly into 4 frames, before getting concatenated into the model's input. Images are downsized so that they have a max pixel count of 524,288.

C Technical Details of the DRPO Algorithm

C.1 Domain-aware Relative Policy Optimization

Group Relative Policy Optimization, or GRPO, is a reinforcement learning algorithm that gained widespread attention following the release of DeepSeek-R1. In contrast to Proximal Policy Optimization, which requires a separate value network to estimate advantages, GRPO computes the advantage $\hat{A}_{(q,i,t)}$ directly from a collection of sampled responses for each prompt. A rollout refers to a single response trajectory generated by the model when presented with a given input. The advantage function measures the relative quality of one response compared to others generated for the same prompt, allowing the optimization procedure to favor higher quality outputs without explicitly estimating expected returns.

For a given prompt q, the algorithm generates a group of rollouts $G_{(q,t)}$ at training iteration t. Each response $o_{(q,i,t)}$ in this group consists of a sequence of tokens

$$o_{(q,i,t)} := o_{(q,i,t):1},\ o_{(q,i,t):2},\ \ldots,\ o_{(q,i,t):n_{o_{(q,i,t)}}},$$

where $n_{o_{(q,i,t)}}$ denotes the total number of tokens in the sequence. Let $r_{(q,i,t)}$ represent the scalar reward assigned to the i-th response. The complete set of rewards for the group is written as $R_{G_{(q,t)}} = \{r_{(q,1,t)}, r_{(q,2,t)}, \ldots, r_{(q,|G_q|,t)}\}$, where $|G_q|$ indicates the number of responses in the group. GRPO standardizes these rewards to have zero mean and unit variance:

$$\hat{A}^{\text{GRPO}}_{(q,i,t)} = \frac{r_{(q,i,t)} - \hat{\mu}_{G_{(q,t)}}}{\hat{\sigma}_{G_{(q,t)}} + \varepsilon}, \tag{8}$$

where $\hat{\mu}_{G_{(q,t)}}$ and $\hat{\sigma}_{G_{(q,t)}}$ represent the empirical mean and standard deviation of the group rewards, and ε is a small constant included for numerical stability.

These advantage estimates are incorporated into a clipped surrogate objective that also includes a per token regularization term based on the Kullback Leibler divergence:

$$\tilde{A}_{(q,i,t):k}(\theta) = \min\left(\varphi_{(q,i,t):k}(\theta)\cdot\hat{A}^{\text{GRPO}}_{(q,i,t)},\ \text{clip}\left(\varphi_{(q,i,t):k}(\theta),\ 1-\varepsilon,\ 1+\varepsilon\right)\cdot\hat{A}^{\text{GRPO}}_{(q,i,t)}\right),$$

$$\varphi_{(q,i,t):k}(\theta) = \frac{\pi_\theta(o_{(q,i,t):k} \mid q,\ o_{(q,i,t):<k})}{\pi_{\theta_{\text{old}}}(o_{(q,i,t):k} \mid q,\ o_{(q,i,t):<k})},$$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathscr{D},\{o_{(q,i,t)}\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|G_{(q,t)}|}\sum_{i=1}^{|G_{(q,t)}|}\frac{1}{n_{o_{(q,i,t)}}}\sum_{k=1}^{n_{o_{(q,i,t)}}}\tilde{A}_{(q,i,t):k}(\theta) - \beta\, D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right].$$

In these expressions, $o_{(q,i,t):<k}$ denotes the token subsequence from position 1 to position $k-1$, and $\mathscr{D}$ represents the training data distribution. The ratio $\varphi_{(q,i,t):k}(\theta)$ captures the importance sampling weight between the current policy π_θ and the previous policy $\pi_{\theta_{\text{old}}}$ at token position k. The coefficient β controls the strength of regularization toward the reference policy π_{ref}.

GRPO performs well when training data exhibits relatively uniform characteristics. However, clinical datasets often contain substantial heterogeneity across different medical domains, imaging modalities, and disease categories. In such settings, domains with more abundant training examples tend to dominate the optimization process, while underrepresented domains contribute minimally to parameter updates. This imbalance can cause the model to perform poorly on rare but clinically significant conditions, while expending excessive computational resources on common cases that the model already handles well.

Domain-aware Relative Policy Optimization. Although GRPO normalizes rewards across responses to the same prompt, it does not account for imbalances across different clinical domains. Domains that appear more frequently in the training data therefore exert disproportionate influence on learning. DRPO extends GRPO by introducing a hierarchical scaling mechanism that explicitly balances contributions from different domains. This correction for cross domain imbalance retains the simplicity of the GRPO formulation while promoting more equitable learning across heterogeneous clinical data.

Hierarchical Scaling Based on Clinical Domain Structure. The central innovation of DRPO is a hierarchical scaling strategy that adaptively balances learning signals according to both domain frequency and task difficulty. This mechanism operates at two levels. At the first level, scaling corrects for the dominance of overrepresented clinical domains. At the second level, scaling adjusts for variations in question difficulty within each domain. The algorithm first clusters questions within each domain based on their reward distributions, treating the set of rollout rewards as a feature vector. A two stage scaling procedure is then applied, first at the cluster level and then at the individual reward level, which emphasizes learning from rare and challenging clinical questions.

Stage 1: Clustering Within Clinical Domains. At each training iteration t, we sample an independent batch of questions and organize them by clinical domain. Let g denote a specific domain, and let $N_{(g,t)}$ represent the number of questions belonging to domain g at iteration t. For each question, we compute rewards across all sampled rollouts and concatenate these into a feature vector. Formally, for domain g, we construct the set of reward vectors

$$\mathscr{H}_g = \{\mathbf{v}_q^g\}_{q=1}^{N_g}, \quad \mathbf{v}_q^g \in \mathbb{R}^{|G_{(q,t)}|},$$

where $\mathbf{v}_q^g$ contains the rollout rewards $R_{G_{(q,t)}}$ for question q, and $N_{(g,t)}$ is the number of questions in domain g at iteration t.

To identify patterns in question difficulty, we apply K-means clustering to these reward vectors separately within each domain:

$$\{\mathbf{C}_{(1,g,t)}, \mathbf{C}_{(2,g,t)}, \ldots, \mathbf{C}_{(k_{(g,t)},g,t)}\} = \text{KMeans}(\mathscr{H}_g, k_{(g,t)}),$$

where $\mathbf{C}_{(c,g,t)}$ denotes the centroid of cluster c in domain g, and $k_{(g,t)}$ is the number of clusters determined automatically using the elbow method.

Stage 2: Hierarchical Temperature Scaling. For each domain and each cluster within that domain, we compute temperature factors that capture both the relative size and average difficulty of questions. Difficulty is estimated using mean reward values, which serve as a proxy for how challenging the questions are. We define inter domain temperature factors $T_{(g,t)}$ and intra domain temperature factors $T_{(c,g,t)}$ as follows:

$$T_{(g,t)} = \max\left(\sqrt{N_{(g,t)}} \cdot \mu_{(g,t)}, \varepsilon\right), \quad T_{(c,g,t)} = \max\left(\sqrt{N_{(c,g,t)}} \cdot \mu_{(c,g,t)}, \varepsilon\right), \tag{9}$$

where $N_{(c,g,t)}$ is the number of questions in cluster c, and $\mu_{(g,t)}$ and $\mu_{(c,g,t)}$ denote the mean rewards for domain g and cluster c respectively.

These temperature factors are inversely applied to the advantage function, so that smaller or more difficult domains and clusters receive proportionally greater weight during training. Rewards are first normalized at the question level as in GRPO, then scaled by the domain and cluster temperatures, and finally multiplied by a regularization factor $m_{(i,t)}$ that prevents outliers from dominating updates:

$$s_{(q,i,t)}^{scaled} = \frac{m_{(i,t)} \cdot s_{(q,i,t)}}{T_{(g,t)} \cdot T_{(c,g,t)}}, \tag{10}$$

where $s_{(q,i,t)} = \frac{r_{(q,i,t)} - \mu_{(q,t)}}{\sigma_{(q,t)} + \varepsilon}$ is the question level normalized reward from GRPO. To restore unit variance, we divide by the standard deviation across the batch:

$$\hat{A}_{(q,i,t)}^{\text{DRPO}} = \frac{s_{(q,i,t)}^{scaled}}{\sigma_{\mathbf{s}_i^{scaled}}}.$$

The DRPO Objective Function. DRPO retains the same objective structure as GRPO, with the modified advantage estimate:

$$\tilde{A}_{(q,i,t):k}(\theta) = \min\left(\varphi_{(q,i,t):k}(\theta) \cdot \hat{A}_{(q,i,t)}^{\text{DRPO}},\ \text{clip}\left(\varphi_{(q,i,t):k}(\theta),\ 1-\varepsilon,\ 1+\varepsilon\right) \cdot \hat{A}_{(q,i,t)}^{\text{DRPO}}\right),$$

$$J_{\text{DRPO}}(\theta) = \mathbb{E}_{q\sim\mathscr{D},\{o_{(q,i,t)}\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|G_{(q,t)}|}\sum_{i=1}^{|G_{(q,t)}|}\frac{1}{n_{o_{(q,i,t)}}}\sum_{k=1}^{n_{o_{(q,i,t)}}}\tilde{A}_{(q,i,t):k}(\theta) - \beta\, D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right],$$

where $\varphi_{(q,i,t):k}(\theta) = \frac{\pi_\theta(o_{(q,i,t):k} | q,\, o_{(q,i,t):<k})}{\pi_{\theta_{\text{old}}}(o_{(q,i,t):k} | q,\, o_{(q,i,t):<k})}$.

Clinical Relevance of DRPO. The hierarchical design of DRPO addresses several challenges that arise when training clinical AI systems:

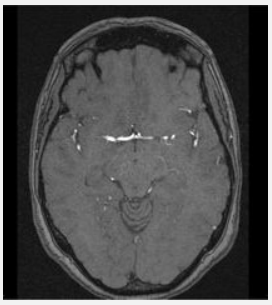
1. **Balanced Learning Across Clinical Domains:** The two layer scaling mechanism directs optimization toward both underrepresented medical domains and challenging clinical questions within each domain. This prevents the model from focusing exclusively on common conditions while neglecting rare but important clinical scenarios, such as unusual disease presentations or atypical imaging findings.
2. **Stable Optimization Properties:** DRPO applies scaling after the initial GRPO normalization step, preserving the property that rewards have zero mean and unit variance within each group of rollouts. This property is important for stable reinforcement learning optimization, as established in prior work on reward normalization[1–3].

3. **Computational Efficiency:** DRPO introduces minimal additional computational cost, with complexity of order $O(n)$ arising primarily from K-means clustering on low dimensional vectors containing typically 5 to 10 elements. This efficiency enables training at scale without requiring auxiliary critic networks, making the approach practical for fine tuning large clinical language models.

D Relevance Annotation of Reasoning Traces

To better understand the reasoning process and help model produce better reasoning in the future, we collaborated with clinicians to annotate the reasoning traces on randomly selected samples in the validation set, which is representative of model's real world performance on unseen samples. We present unfiltered, original model reasoning traces here with color-coded highlights indicating relevance. Phrases marked in yellow represent information of low clinical relevance, pink highlights indicate high clinical relevance, and aqua denotes content judged to have no relevance to the diagnosis.

Question: Above is a brain MRI scan of a patient. What type of tumor, if any, is present in this brain MRI scan?

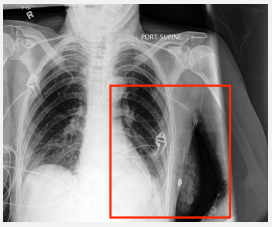


Answer with one of the following: No Tumor, Glioma Tumor, Meningioma Tumor, Pituitary Tumor
Ground Truth Label: No Tumor
Model's Prediction: No Tumor

Reasoning: To analyze the brain MRI scan, we should consider the following steps:

1. **Understanding MRI Scans:**
   - **MRI (Magnetic Resonance Imaging)** is a diagnostic tool that uses a powerful magnet, radio waves, and a computer to create detailed images of organs and structures inside the body, including the brain.
   - It is particularly useful for identifying tumors because it can show differences in tissue density, which can indicate whether a mass is solid, liquid, or blood-filled, and whether it is benign or malignant.
2. **Identify Key Features of a Tumor on MRI:**
   - **Hypointense/Isointense on T1:**
     - Most normal brain tissue appears dark on a T1-weighted MRI, so a tumor that is the same color as normal brain tissue is called "isointense" or "hypointense."
     - This means it is often harder to spot on a T1 image, but can be seen on the contrast-enhanced images.
     - A tumor that is the same color as brain tissue on a T1-weighted MRI is often a sign of a glioma, which is a type of brain tumor that originates from glial cells in the brain.
   - **Hyperintense on T2:**
     - A tumor that appears bright on a T2-weighted MRI is called "hyperintense."
     - This means it is often easier to see on a T2 image, but is a less specific sign.
     - It can be seen in many brain tumors, including gliomas, meningiomas, and pituitary tumors.
     - The brightness on T2 images is often associated with edema, which is a sign of inflammation and swelling around the tumor, which is a common feature of many types of tumors.
     - The image provided is a T2 weighted MRI, which usually makes tumors appear bright, and in this case, the tumor appears hyperintense, consistent with a typical appearance of a meningioma on MRI. Meningiomas are often hyperintense on T2 images and hypointense on T1 images, which is what we see here. They are usually benign but can be quite large and cause symptoms. However, without further context, it is impossible to determine definitively the type of tumor. The MRI is simply a tool, and a definitive diagnosis often requires a combination of imaging, a patient's medical history, and other tests such as a biopsy.

Question: Above is a chest X-ray image of a patient. What is the diagnosis of the patient in the X-ray image?

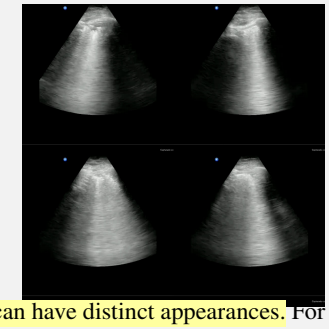


Answer with one or multiple phrases from the following: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices
Ground Truth Label: Support Devices
Model's Prediction: No Findings

Reasoning to annotate: To analyze the chest x-ray image, we need to consider the following steps:
1. **Examine the Image for Anomalies:**
    - **Size and Shape of the Heart and Mediastinum:** The mediastinum is the area in the chest that contains the heart, great vessels, trachea, and esophagus. In this chest x-ray, the mediastinal area appears to be enlarged, which is a strong indicator of pathologies such as cardiomegaly (an enlarged heart) or other mediastinal pathologies, but it's essential to understand that this is an interpretation and should be confirmed by a medical professional.
    - **Lung Opacity:** The lung fields are evaluated for any signs of increased density, which can indicate fluid, air, or tissue abnormalities. The x-ray shows increased density in the lung fields, which can suggest consolidation, such as from pneumonia, or other pathology. The appearance of the lung fields is consistent with signs of fluid, which is a sign of pleural effusion.
    - **Support Devices:** The x-ray shows a pacemaker, which is a medical device used to regulate the heartbeat. This is a normal finding in some patients and does not indicate pathology.
2. **Consider the X-ray Protocol and Patient Information:**
    - The patient's position (supine) is noted. This position is often used in chest x-rays to ensure that the patient is in a comfortable position and to ensure that the image is clear and useful for interpretation. The radiographic markers (e.g., the "P" and "R" signs) are standard for x-rays to indicate the position of the patient and to ensure that the image is correctly oriented.
3. **Analysis of the Image for Any Indications of Pathology:**
    - **Pleural Effusion:** The increased density in the lung fields is indicative of pleural effusion, which is a buildup of fluid in the pleural space, the area between the lungs and the chest wall. This is a common sign of heart failure, liver disease, or malignancy.
    - **Cardiomegaly:** The enlarged heart size, though not definitive without further clinical context, is suggested by the mediastinal widening, though the definitive interpretation is that this is more about the mediastinum and not a direct indicator of cardiomegaly.

Based on the analysis, the most likely findings in this x-ray image are suggestive of:
- Pleural Effusion due to the increased density seen in the lung fields.
- No Finding for other abnormalities, given that there are no signs of fractures, pneumothorax, or pneumonia.

Question: Above is a lung ultrasound video. What is the diagnosis based on this lung ultrasound?

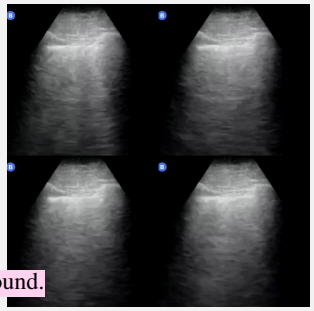


Answer with one word from the following options: covid, pneumonia, normal
Ground Truth Label: covid
Model's Prediction: pneumonia

Reasoning to annotate: To analyze the image and determine the reasoning process for making a diagnosis, we follow these steps:

1. **Examine the Image Type and Purpose**: The image is a lung ultrasound, which is a diagnostic tool used to assess lung parenchyma (the part of the lung that contains air and is responsible for gas exchange) for signs of abnormalities. Ultrasound is particularly useful in the early stages of respiratory disease, as it can detect changes in lung tissue that may be indicative of various conditions.
2. **Understanding Ultrasound Features**: In ultrasound, different tissue types and pathologies can have distinct appearances. For example:
    - **Hyperechoic (bright) areas** often indicate a pathological process, such as air, fluid, or solid tissue. In lung ultrasound, hyperechoic areas might suggest consolidation, which is often associated with pneumonia.
    - **Hypoechoic (dark) areas** are usually indicative of normal lung tissue, but in some cases, they could also be associated with certain pathologies.
    - **Echogenicity patterns** can vary in different lung pathologies, and these patterns can help in identifying the nature of lung pathology. For example, the image shows a relatively uniform, hyperechoic pattern, which is often seen in lung ultrasound when it is used to identify areas of consolidation, a hallmark of pneumonia.
3. **Clinical Correlation**: It is essential to consider the patient's clinical context, such as symptoms, medical history, and other diagnostic tools used in conjunction with the ultrasound. However, since the image is just an ultrasound, we must rely on the image's characteristics alone. The image shows a single ultrasound frame, and often, a series of ultrasound images are analyzed over time to get a comprehensive view of lung involvement, but for this exercise, we are only analyzing one image.
4. **Interpretation of Image**: In this case, the image shows a pattern that is consistent with lung ultrasound findings that suggest pneumonia. The hyperechoic areas are characteristic of consolidation, which is a sign of fluid or air accumulation in the lungs, typically associated with pneumonia. This is further supported by the fact that lung ultrasound is a sensitive tool for identifying pneumonia in its early stages, often before radiographic changes are visible.

Question: Above is a lung ultrasound video. What is the diagnosis based on this lung ultrasound?



Answer with one word from the following options: covid, pneumonia, normal
Ground Truth Label: covid
Model's Prediction: pneumonia

Reasoning to annotate: To analyze the image and determine the reasoning process for understanding what the image might suggest, follow these steps:
1. **Examine the Image Type:** The image is a lung ultrasound, which is a diagnostic imaging technique used in medicine to visualize the lung parenchyma. Ultrasound is often used in emergency settings for quick, non-invasive assessment, especially in cases where quick decision-making is needed. Ultrasound images are grayscale, with varying shades of gray representing different tissue densities. The image shown here is a grayscale ultrasound image, which is typical for ultrasound.
2. **Analyze the Image Features:**
   - **Grayscale Intensity:** The image is grayscale, with varying shades of gray indicating different tissue densities. In ultrasound, the brightness of the image can be indicative of different pathologies. For example:
   - **Hyperechoic (brighter than normal tissue):** This often indicates air, which is why ultrasound is used to visualize the chest, as air is easily detected. In this image, the lesion appears hyperechoic, meaning it reflects more ultrasound waves than the surrounding tissue, which is a characteristic sign that can suggest the presence of fluid or air, which is often associated with disease such as pneumonia.
   - **Hypoechoic (darker than normal tissue):** This indicates less reflective tissue and is often associated with soft tissues, such as tumors, or areas of fluid accumulation. The lesion in this image is not hyperechoic, so it is unlikely to suggest air, but rather, it suggests a lesion that is darker than the surrounding lung tissue, which could be indicative of a pathological condition.
   - **Shape, Size, and Border:** The lesion in the image is well-defined, meaning it has a clear border, which is a feature that can be associated with malignancy or inflammation. The shape is mostly round, which is common in benign lesions. However, the exact nature of the lesion would need further investigation to determine if it is benign or malignant. The size and location are critical, but these cannot be determined from a single image.

3. **Clinical Context:** Ultrasound is often used in the context of acute respiratory distress, and it is frequently used to identify signs of pneumonia, which can be associated with various signs on ultrasound such as consolidation, pleural effusion, or air. The image shows a lesion, which may suggest pathology, but it is crucial to understand the clinical context, such as the patient's history, symptoms, and other clinical signs, to make a full diagnosis.
Based on the reasoning process above, the analysis of this ultrasound image suggests that it is showing signs that might be indicative of pneumonia. The hyperechoic nature of the lesion is often associated with consolidation, a sign that can occur in pneumonia. However, it's important to remember that ultrasound is a tool and can only suggest possibilities; the final diagnosis is typically confirmed by further tests, such as chest X-rays, CT scans, or other medical procedures.

Question: Above is a chest X-ray image of a patient. What is the diagnosis of the patient in the X-ray image?

Answer with one or multiple phrases from the following: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices
Ground Truth Label: Pneumonia
Model's Prediction: Pneumonia

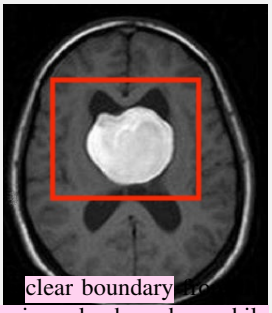
Reasoning to annotate: To analyze the chest x-ray image, we need to understand the reasoning process for identifying signs that may suggest a particular diagnosis. The reasoning process involves observing the following key features:

2. **Lung Opacity:**
   - The density of the lung fields is evaluated. Normally, lung tissue is very aerated and appears very light on the x-ray, almost completely black. If there is any increased opacity in the lung fields, it may indicate fluid, blood, or abnormal tissue, such as consolidation, atelectasis, or consolidation, which are suggestive of pneumonia. In this image, the lung fields appear darker than normal, which is a sign of increased opacity. This could indicate a number of pathologies, such as consolidation, which is often associated with pneumonia. Atelectasis, which is the collapse of part of the lung, can also appear as a darker opacity on x-ray and is a common cause of increased lung opacity.
3. **Pleural Effusion:**
   - This is where there is an abnormal accumulation of fluid in the pleural space, which can be seen on x-rays as a smooth, curved opacity at the base of the lung, often associated with lung collapse.
   - This isn't visible in this image, as there's no indication of a smooth, curved opacity.
4. **Pneumothorax:**
   - This is where air is in the pleural space, and it is usually seen as a black, triangular area on x-rays, because air is very radiolucent and doesn't allow any x-rays to pass through.
   - This image does not have any signs of pneumothorax, such as a triangular black area.
5. **Pleural Other:**
   - This can include other pathologies that affect the pleura, such as malignancy, infection, or inflammation. However, this image does not show signs of malignancy or inflammation, so it is unlikely to be the primary focus.
6. **No Finding:**
   - This means that the x-ray shows no obvious signs of pathology, but this is unlikely in this image, as there are signs that suggest pathology.
7. **Lung Lesion:**
   - This is a general term for any abnormality in the lung, which can be a sign of pathology. The darker, more opaque areas in the lungs are suggestive of a lung lesion, which could be caused by a variety of reasons including pneumonia, atelectasis, or consolidation. In this image, the areas of increased opacity in the lung fields are suggestive of a lung lesion.

Based on the reasoning process, the most plausible findings in this image are those that suggest increased lung opacity, which are suggestive of pathology. The reasoning process points towards lung lesions, which are often associated with pneumonia, atelectasis, or consolidation. The signs of increased lung opacity are indicative of a pathology that could be pneumonia, which is a common cause of increased lung opacity in x-rays.

Question: Above is a brain MRI scan. Is there a tumor present in this brain MRI scan?

Answer with one of the following: No Tumor, Has Tumor
Ground Truth Label: Has Tumor
Model's Prediction: Has Tumor

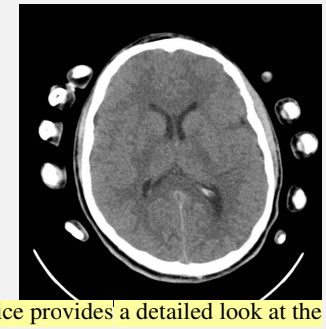


Reasoning to annotate: To analyze the image, let's use the following reasoning steps:

1. **Identify the Type of Image:** The image is a brain MRI scan, which is a medical imaging technique used to visualize detailed cross-sectional images of the brain. MRI is particularly useful for detecting tumors because it provides high contrast between different types of tissues, including tumors, which can be seen as distinct from normal brain tissue.
2. **Analyze the Image Features:**

- **Shape and Size:** The tumor appears as a round, well-circumscribed lesion, meaning it has clear boundary surrounding brain tissue. In MRI images, tumors are often characterized as having a distinct, often irregular, boundary, while normal brain tissue has a smooth, well-defined edge. In this case, the tumor is clearly distinct from the surrounding brain tissue, which is a strong indicator of a lesion that could be a tumor.
- **Signal Intensity:** The signal intensity refers to how bright or dark the tissue appears on the MRI image. Tumors can vary in signal intensity, but often, they appear brighter than normal brain tissue on a T2-weighted MRI (which is the standard MRI sequence used for brain imaging). This is because tumors have different tissue properties, such as increased blood flow, which can lead to a higher signal intensity on MRI. This tumor is indeed bright on this image, suggesting it is a high-intensity lesion, which is often associated with tumors, but this alone is not definitive for a diagnosis.
- **Location and Location:** The tumor is located in the brain, and in this case, it's in a region where tumors are common. The location and size of the lesion are important factors in determining the potential pathology, but they are not definitive on their own. The location in this image is consistent with where tumors commonly appear in the brain, and the size is indicative of a lesion that could be a tumor.
- **Clinical Context:** While this reasoning relies mostly on the image itself, the context (clinical history of the patient, symptoms, and any other imaging studies) would be considered by a medical professional to make a definitive diagnosis. This context is often essential for a precise understanding and interpretation of the image.

Question: Above is a brain CT scan slice. Is there any hemorrhage present in this CT scan slice?



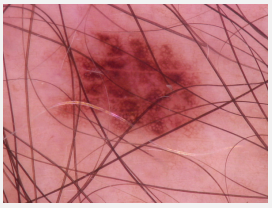
Answer with one of the following: No Hemorrhage, Has Hemorrhage
Ground Truth Label: No Hemorrhage
Model's Prediction: No Hemorrhage

Reasoning to annotate: To analyze the image, we will go through a series of steps to understand what we're looking at and what it might tell us about the presence of hemorrhage.

1. **Understanding CT Scans:** CT (computed tomography) scans use X-rays and computer processing to create detailed images of the body's internal structures. They are particularly useful for diagnosing brain injuries because they can show changes in tissue density, which can indicate bleeding, tumors, or other pathologies. The color and density of the tissue on a CT scan can often give clues about what is happening.
2. **Anatomy of a CT Scan Image:** A CT scan is a cross-sectional view of the body, and each slice provides a detailed look at the brain at that level. The image is typically grayscale, with different shades representing different densities. The brain tissue usually appears as a medium gray, while air or gas appears black, and bone, which is very dense, appears white.
3. **Color Coding on CT Scans:** The way tissues appear on CT scans can vary depending on the type of contrast used and the type of scan (e.g., CT with contrast, CT without contrast). In a standard CT scan, brain tissue usually appears as a medium gray. If there is bleeding, the blood can appear in different shades on CT scans, depending on the type and age of the hemorrhage. Typically, blood is denser than brain tissue and appears white on a non-contrast CT scan, while on a contrast-enhanced CT scan, it can appear darker (hypodense) or have a different color depending on its age and location.
4. **Symptoms of Hemorrhage on CT:** On a CT scan, signs of hemorrhage include:
    - **Hypodense (darker than normal brain tissue) on non-contrast CT**: This is the most common appearance for acute hemorrhage. It happens when blood clots and does not allow contrast to pass through.
    - **Hyperdense (whiter than normal brain tissue) on non-contrast CT**: This is seen with older blood, where the blood has already clotted and is more dense than normal brain tissue.
    - **Contrast enhancement**: If a contrast medium is used, it can help highlight the blood clot, making it stand out even more on the scan.
5. **Analyzing the Image:**
    - The image is in grayscale, and in this case, the brain tissue appears in a medium gray shade, which is normal for brain tissue in a CT scan.
    - There is no evidence of white areas (hyperdense), which would suggest older blood clots or contrast enhancement that would indicate bleeding. The entire brain tissue appears uniformly gray, without any obvious bright or dark areas that would suggest hemorrhage.

Based on these considerations, we are looking for signs that would indicate hemorrhage, but there are none in this image. The tissue density is consistent with normal brain tissue, and there is no evidence of blood clots or other signs of hemorrhage. The image is consistent with a brain CT without hemorrhage.

Question: Above is a dermoscopy image of a patient. What is the diagnosis of the patient in the dermoscopy image?



Answer with one word from the following: Melanoma, Nevus, Basal Cell Carcinoma, AKIEC, Other
Ground Truth Label: Nevus
Model's Prediction: Nevus

Reasoning to annotate: To analyze the image, let's go through the reasoning process step by step:

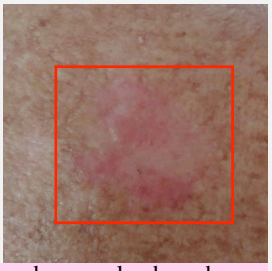
1. **Examine the Image:**

- The image shows a close-up of skin with a lesion. The lesion appears to have a somewhat uneven, irregular, and slightly raised surface, which can be indicative of various skin conditions. The coloration is mostly a dark brown, which could be a sign of a pigment-rich lesion, such as a mole or a nevus. The border of the lesion appears somewhat uneven, which is a characteristic feature of many nevi, but this is not definitive without further context.

2. **Consider the Clinical Context:**

- In dermatology, dermoscopy is often used to evaluate skin lesions for signs that may suggest malignancy. The image provided does not have any clinical history or patient data, but based on the visual clues, we can consider the following points:
  - **Melanoma** typically shows a variety of features that can be indicative of malignancy, such as a changing color, irregular border, or asymmetry. The lesion in this image does not show any obvious signs of malignancy like ulceration, bleeding, or bleeding into the lesion, which are often associated with melanoma. However, these signs are not always present, and a definitive diagnosis requires a biopsy.
  - **Nevus** (or mole) is a common benign growth of the skin. These are often characterized by their uniformity, color, and border. In this image, the lesion's color is consistent, and its border is somewhat irregular, but this is not a definitive sign for malignancy. The lesion could be a nevus, but we cannot make a certain diagnosis without further clinical context and examination.
  - **Basal Cell Carcinoma** (BCC) is usually a non-melanoma skin cancer. BCCs are often erythematous, ulcerated, and can be very vascular. The lesion in this image lacks these characteristics, so it is less likely to be BCC. However, it is still a possibility, but it is less likely based on the image alone.
  - **AKIEC** (Actinic Keratoses) are precancerous skin lesions that are usually caused by chronic sun exposure. They often have a rough, scaly, or warty appearance, and they are usually a light tan, brown, or red. This lesion does not have the characteristics of AKIEC, and the image does not suggest a precancerous lesion.
  - **Other**: There are other possibilities such as a vascular lesion, which could be due to various causes such as angiomas, which are benign. However, the image does not provide enough information to make a definitive diagnosis, and further clinical examination and histopathology are required.

Question: Above is a clinical image of a patient. What is the diagnosis of the patient in the clinical image?



Answer with one word from the following: Melanoma, Nevus, Basal Cell Carcinoma, AKIEC, Other
Ground Truth Label: Basal Cell Carcinoma
Model's Prediction: Basal Cell Carcinoma

Reasoning to annotate: To analyze the image, let's go through the reasoning process step by step:

1. **Observation of the Image:**
- **Color and Texture:** The skin lesion in the image is pinkish in color, which can be indicative of various skin conditions. The color suggests it's not a dark mole (nevus) or a dark, raised, and unevenly shaped mass (melanoma). The color is relatively uniform, but the pink hue is a bit unusual for normal skin, indicating something might be amiss.
- **Shape and Size:** The lesion appears as a slightly raised, pinkish, and slightly elevated area, with a somewhat indistinct border. The uneven border is a critical feature that could suggest a basal cell carcinoma, as these often have a more irregular border, while melanoma is known for having a jagged, uneven, or indistinct border. Nevus usually have a more sharply defined border and are often flat, while AKIEC (Actinic Keratosis) lesions have a rough, scaly surface, which is more consistent with the texture in this image. The lesion's size and shape are also consistent with basal cell carcinoma, which is more likely to present as a raised, pink, or flesh-colored lesion on the skin, especially on sun-exposed areas.
- **Skin Lesion Location and Context:** The lesion is on the skin, which is a common location for basal cell carcinomas, which are the most common type of skin cancer. They often appear on areas most exposed to the sun, such as the face, ears, neck, and back, which aligns with the location of the lesion in this image.
- **Clinical Context:** Basal cell carcinomas are typically painless and are often associated with chronic sun exposure. They are usually benign, but they can be removed to prevent potential malignancy. In this context, the lesion's color, shape, and location are consistent with basal cell carcinoma. It is important to remember that only a dermatologist can make a definitive diagnosis, but this analysis is based on common clinical patterns.

E Q&A Pair Examples of Johns Hopkins Hospital Data

Example 1: Length of Stay Prediction
Patient History: A 17-year-old male (pediatric, not Hispanic) with malignant neoplasm of overlapping sites of brain (ICD-10: C71.8, NEOPLASMS category). Low comorbidity burden (1 diagnosis), no ICU admission. Elective admission to Radiation Oncology at 2018-06-26 07:57:00.
Question: How long will the patient stay in the hospital? **Options:** A. Short stay (0-2 days), B. Moderate stay (3-7 days), C. Extended stay (8-14 days), D. Long-term stay (>14 days).
Ground Truth: A - Short stay (0-2 days), Actual LOS: 0.67 days (16.03 hours)
Model Prediction: A
Explanation: The patient is a 17-year-old male with a primary diagnosis of a malignant neoplasm of overlapping sites of the brain, which is a serious condition that often requires hospitalization for treatment. The fact that this is an elective admission to the Radiation Oncology department suggests that the patient is likely receiving a planned course of treatment, which typically involves multiple sessions of radiation therapy. Elective admissions for radiation therapy are often associated with shorter hospital stays, usually ranging from 0 to 2 days, as the patient may be admitted for the initial setup and then discharged to receive the treatment at home or in a clinic.
Result: ✓ CORRECT

Example 2: Length of Stay Prediction (ESRD)
Patient History: A 19-year-old female (young adult, not Hispanic) with ESRD on hemodialysis (ICD-10: N18.6, GENITOURINARY SYSTEM category). Low comorbidity burden (1 diagnosis), no ICU admission. Elective admission to Pediatric Dialysis at 2020-12-04 06:31:00.
Question: How long will the patient stay in the hospital? **Options:** A. Short stay (0-2 days), B. Moderate stay (3-7 days), C. Extended

Table 1. Patient Demographics Data Structure in ACCM PMAP Dataset

Data Element	Description/Values
Patient Identifiers	
osler_id	Patient unique identifier
emrn	Enterprise MRN
jhhmrn, bmcmrn, hcgmrn, smhmrn, shmrn	Facility-specific MRNs
Basic Demographics	
Name fields	lastname, firstname, middlename, namesuffix
birth_date	Date of birth
age	Derived from DOB
Vital Status	
pat_status	Alive or Deceased
death_date	Date of Death (if applicable)
Gender	
gender	Female, Male, Nonbinary, Other, Unknown
genderabbr	F, M, BN, O, U
Ethnicity	
ethnic_group	Hispanic, Not Hispanic, Pt Refused, Unknown
Race (Multiple Categories Possible)	
first_race	Primary documented race
racew, raceb, racei, racea, racep, raceo	White, Black, Am Indian, Asian, Pacific Islander, Other ('Y' or Null)
racetwo	Two or more races ('Y' or Null)
racedec	Declined race ('Y' or Null)
raceu	Unknown race ('Y' or Null)
Geographic Information	
Address	address1, address2, city, stateabbr, county, country, zip-code
Geocoding	latitude, longitude, census tract (FIPS format)
Social Demographics	
marital_status	Declined, Divorced, Separated, Married, Other, Significant Other, Single, Unknown, Widowed
language	Patient language
employment_status	Disabled, Full Time, Never Worked, Not Employed, Military, Part Time, Retired, Self Employed, Student, Unknown

stay (8-14 days), D. Long-term stay (>14 days).
Ground Truth: A - Short stay (0-2 days), Actual LOS: 0.73 days (17.47 hours)
Model Prediction: D
Explanation: The patient is on hemodialysis, which is a long-term treatment for end-stage renal disease (ESRD). Hemodialysis typically requires a patient to be in the hospital for several hours each day, often several days a week, and can extend over many months or years. Given that the patient is on hemodialysis, it is reasonable to conclude that they will have a long-term stay in the hospital, which is defined as more than 14 days.
Result: × INCORRECT

Example 3: Mechanical Ventilation Duration
Patient History: A 36-year-old female (young adult, not Hispanic) with cervical cerclage suture present, second trimester (ICD-10: O34.32). No ICU admission, no ventilation support documented. Elective admission to Maternal and Fetal Medicine at 2015-07-21 11:00:00.

Question: What is the total duration of mechanical ventilation (invasive ventilation only, level 6-7: Conventional Mechanical Ventilation or High-Frequency Oscillatory/Jet Ventilation) for this patient during the encounter? **Options:** A. No invasive ventilation, B. Less than 24 hours, C. 24-72 hours, D. Greater than 72 hours.
Ground Truth: A - No invasive ventilation, Actual Invasive Ventilation: 0.00 hours (0 episodes)
Model Prediction: A
Explanation: The patient's history indicates no ventilation support was documented, and there is no mention of any mechanical ventilation, invasive or otherwise, during the hospital encounter. The admission was elective and related to maternal-fetal medicine, with no indication of respiratory failure or need for ventilatory support.
Result: ✓ CORRECT

Example 4: Sepsis Classification
Patient History: A 71-year-old female (elderly, not Hispanic) with hypotension (ICD-10: I95.9, CIRCULATORY SYSTEM category). No infection-related diagnoses. Organ dysfunction: Cardiovascular hypotension (I95.9) and shock circulatory (R57.9). ICU admission (hoco_icu). Emergency admission via ED at 2014-01-12 21:23:00, Surgery department, Internal Medicine service.
Question: Did this patient receive a diagnosis of sepsis, severe sepsis, or septic shock during this encounter? **Options:** A. No sepsis diagnosis, B. Sepsis (ICD-10 A40-A41, without organ dysfunction), C. Severe sepsis (ICD-10 R65.20), D. Septic shock (ICD-10 R65.21).
Ground Truth: A - No sepsis diagnosis
Model Prediction: A
Explanation: The patient's documented diagnoses are hypotension (I95.9) and shock circulatory (R57.9), which are non-infectious and not specifically linked to sepsis. There are no infection-related diagnoses, and no ICD-10 codes for sepsis, severe sepsis, or septic shock are present. Therefore, sepsis was not diagnosed during this encounter.
Result: ✓ CORRECT

Example 5: Septic Shock Misclassification
Patient History: A 64-year-old male (middle-aged, not Hispanic) with AKI (ICD-10: N17.9, GENITOURINARY SYSTEM category). Infection-related diagnoses: acute pyelonephritis (N10), ileus due to infection (K56.7), centrilobular emphysema (J43.2), UTI (N39.0), septic shock due to UTI (A41.9). Multiple organ dysfunctions: respiratory failure (J96.01, J96.21), cardiovascular shock (R57.9), hypotension (I95.9), renal AKI (N17.9). ICU admission (bmc_cicu). Emergency admission via ED at 2024-02-18 12:10:00, Collaborative Internal Med service.
Question: Did this patient receive a diagnosis of sepsis, severe sepsis, or septic shock during this encounter? **Options:** A. No sepsis diagnosis, B. Sepsis (ICD-10 A40-A41, without organ dysfunction), C. Severe sepsis (ICD-10 R65.20), D. Septic shock (ICD-10 R65.21).
Ground Truth: B - Sepsis (ICD-10 A40-A41, without organ dysfunction), Sepsis Codes: A41.9 (Septic shock due to urinary tract infection)
Model Prediction: D
Explanation: The patient was diagnosed with "Septic shock due to urinary tract infection" (ICD-10: A41.9), which directly corresponds to septic shock. While A41.9 is a specific cause code, the clinical manifestation of septic shock is captured by ICD-10 R65.21, which is the appropriate code for septic shock in the context of this encounter.
Result: × INCORRECT (Model escalated A41.9 to R65.21, misunderstanding ICD-10 coding hierarchy)

Example 6: 30-Day Mortality - Pediatric Minor Injury
Patient History: A 4-year-old male (pediatric, not Hispanic) with contusion of forehead (ICD-10: S00.83XA). Low comorbidity burden (1 diagnosis), no ICU admission, no mechanical ventilation, no sepsis. Emergency admission via ED to Pediatric Emergency Medicine at 2017-10-30 15:33:00, discharged home at 2017-10-30 16:15:00 (LOS: 0.0 days). No mortality risk factors identified.
Question: Did this patient die within 30 days of hospital discharge from this encounter? **Options:** A. Alive at 30 days post-discharge, B. Died within 30 days of discharge, C. Unknown - insufficient follow-up data, D. Still hospitalized at 30 days.
Ground Truth: A - Alive at 30 days post-discharge
Model Prediction: A
Explanation: The patient was a 4-year-old male with a minor forehead contusion, no ICU admission, no sepsis, and no significant comorbidities. He was discharged home after a very brief hospital stay and had no identified mortality risk factors. Given the low

severity of illness and absence of complications, it is highly unlikely he died within 30 days of discharge.
Result: ✓ CORRECT

Example 7: 30-Day Mortality - Dissecting Aortic Aneurysm
Patient History: An 84-year-old male (elderly, ethnicity unknown) with dissecting aortic aneurysm, abdominal (ICD-10: I71.02) and colitis - presumed infectious origin (K52.9). Low comorbidity burden (2 diagnoses), no ICU admission, no mechanical ventilation, no sepsis. Emergency admission via ED to Emergency Medicine at 2013-08-16 21:06:00, discharged at 2013-08-17 03:45:00 (LOS: 0.3 days). Discharge disposition: Other/Unknown. Mortality risk factor: advanced age (84 years).
Question: Did this patient die within 30 days of hospital discharge from this encounter? **Options:** A. Alive at 30 days post-discharge, B. Died within 30 days of discharge, C. Unknown - insufficient follow-up data, D. Still hospitalized at 30 days.
Ground Truth: B - Died within 30 days of discharge, Death Date: 2013-08-31 12:45:00
Model Prediction: A
Explanation: The patient survived to discharge but was 84 years old with a dissecting aortic aneurysm and colitis—conditions that carry significant post-discharge mortality risk. However, the record explicitly states "In-Hospital Mortality: No" and provides no evidence of death within 30 days, nor any follow-up data to suggest otherwise. Thus, based on available information, the patient is presumed alive at 30 days post-discharge.
Result: × INCORRECT (Model failed to recognize high mortality risk of dissecting aortic aneurysm in elderly patient; patient died 14 days post-discharge)

F Q&A Pair Examples of MIMIC-IV Data

Open-ended, list of diagnoses (T1)
Input: Below is a history of a patient: Patient is a 55.0-year-old F.
Procedures performed: Injection or infusion of biological response modifier [BRM] as an antineoplastic agent (ICD code: 9928).
Lab results: Anion Gap: 11.0 mEq/L (2194-11-15 04:45:00); Bicarbonate: 28.0 mEq/L (2194-11-15 04:45:00); Calcium, Total: 8.6 mg/dL (2194-11-15 04:45:00); Chloride: 108.0 mEq/L (2194-11-15 04:45:00); Cholesterol, Total: 299.0 mg/dL (2194-11-15 04:45:00); Creatinine: 0.6 mg/dL (2194-11-15 04:45:00); Magnesium: 2.0 mg/dL (2194-11-15 04:45:00); Phosphate: 3.4 mg/dL (2194-11-15 04:45:00); Potassium: 3.8 mEq/L (2194-11-15 04:45:00); Sodium: 143.0 mEq/L (2194-11-15 04:45:00); Urea Nitrogen: 9.0 mg/dL (2194-11-15 04.
Admitted on 2194-11-14 17:13:00 via EU OBSERVATION from EMERGENCY ROOM.
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission.
The answer should start with “All diagnoses from this admission: ”
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within tags.
Ground Truth: All diagnoses from this admission: E87.6-Hypokalemia; E86.1-Hypovolemia; G35-Multiple sclerosis; R55-Syncope and collapse; R63.4-Abnormal weight loss; E78.5-Hyperlipidemia, unspecified; E87.3-Alkalosis; F41.9-Anxiety disorder, unspecified; F32.9-Major depressive disorder, single episode, unspecified; N02.2-Recurrent and persistent hematuria with diffuse membranous glomerulonephritis; D72.820-Lymphocytosis (symptomatic); E03.9-Hypothyroidism, unspecified; I10-Essential (primary) hypertension.
Model Output: I need to analyze the patient’s history to determine the most likely diagnoses for their current admission. The patient is a 55-year-old female who was admitted via emergency observation after receiving an injection or infusion of a biological response modifier (BRM) as an antineoplastic agent. This suggests she is undergoing cancer treatment.
Looking at the lab results, I see that her electrolytes are mostly within normal limits, with the exception of calcium (8.6 mg/dL, which is low-normal) and magnesium (2.0 mg/dL, which is high-normal). Her anion gap is normal (11.0 mEq/L), bicarbonate is normal (28.0 mEq/L), and her renal function (creatinine 0.6 mg/dL, BUN 9.0 mg/dL) is good. Her cholesterol is elevated (299.0 mg/dL), which is common in patients with cancer or those on certain treatments.
The X-ray images show a normal chest with no obvious signs of infection, tumor, or fluid accumulation. The ECG is not provided, but since it’s not mentioned, I will focus on the other data.
Given that the patient is receiving antineoplastic therapy, the most likely diagnoses would be related to cancer or its treatment. The most common cancers in women over 55 include breast cancer, lung cancer, and colorectal cancer. However, since the X-ray is normal, it’s less likely to be a lung cancer. The patient’s age and gender make breast cancer a likely diagnosis.
Another possibility is that the patient is being treated for a hematologic malignancy, such as lymphoma or leukemia, which could explain the use of a BRM.
The elevated cholesterol could be due to the cancer or the treatment, but it’s not a specific diagnosis.
The most likely diagnoses for this patient are:
Breast cancer (most common in women over 55)

Lymphoma (common in patients receiving BRMs)
Colorectal cancer (common in patients over 55)
These are the most likely diagnoses based on the patient's age, gender, and treatment history.

All diagnoses from this admission: Breast cancer, Lymphoma, Colorectal cancer

Open-ended, single diagnosis (T2)

Input: Below is a history of a patient: Patient is a 81.0-year-old F.
[Summary of prior encounters]
Procedures performed: Drainage of Right Pleural Cavity with Drainage Device, Percutaneous Approach (ICD code: 0W9930Z); Performance of Urinary Filtration, Intermittent, Less than 6 Hours Per Day (ICD code: 5A1D70Z); Control Bleeding in Gastrointestinal Tract, Via Natural or Artificial Opening Endoscopic (ICD code: 0W3P8ZZ); Performance of Urinary Filtration, Intermittent, Less than 6 Hours Per Day (ICD code: 5A1D70Z); Inspection of Upper Intestinal Tract, Via Natural or Artificial Opening (ICD code: 0DJ07ZZ); Destruc.
Lab results: Hematocrit: 30.0 % (2187-11-01 06:45:00); Hemoglobin: 9.6 g/dL (2187-11-01 06:45:00); MCH: 29.2 pg (2187-11-01 06:45:00); MCHC: 32.0 g/dL (2187-11-01 06:45:00); MCV: 91.0 fL (2187-11-01 06:45:00); Platelet Count: 339.0 K/uL (2187-11-01 06:45:00); RDW: 15.2 % (2187-11-01 06:45:00); Red Blood Cells: 3.29 m/uL (2187-11-01 06:45:00); White Blood Cells: 6.8 K/uL (2187-11-01 06:45:00); RDW-SD: 50.4 fL (2187-11-01 06:45:00); Alanine Aminotransferase (ALT): 29.0 IU/L (2187-10-28 02:09:00); Alkaline Phos.
Chart measurements: Glucose finger stick (range 70-100): 149.0 None (2187-10-29 22:00:00); Admission Weight (Kg): 60.1 kg (2187-10-27 06:30:00); Insulin pump: 0.0 None (2187-10-27 08:00:00); Home TF: 0.0 None (2187-10-27 08:00:00); Pressure Ulcer Present: 0.0 None (2187-10-27 08:00:00); 20 Gauge placed in outside facility: 0.0 None (2187-10-27 16:00:00); 20 Gauge Dressing Occlusive: 1.0 None (2187-10-27 16:00:00); 20 Gauge placed in the field: 0.0 None (2187-10-27 16:00:00); Heart Rate: 61.0 bpm (2187-10-27 18:00:0.
Admitted on 2187-10-27 06:25:00 via EW EMER. from TRANSFER FROM HOSPITAL.
What is the projected diagnosis of the patient at this time?
The answer should start with "The primary diagnosis of this admission is " You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within tags. The final answer MUST BE put in \boxed{}.
Ground Truth: The primary diagnosis of this admission is K31.811-Angiodysplasia of stomach and duodenum with bleeding. The primary diagnosis of this admission is categorized as ICD Chapter 11: Diseases of the digestive system (K00-K95). To be more specific, it is within the Diseases of esophagus, stomach and duodenum (K20-K31) block.
Model Output: I need to determine the projected diagnosis for this patient at this time. The patient is an 81-year-old female admitted via emergency with a history of multiple chronic conditions including iron deficiency anemia, hypothyroidism, type 2 diabetes with complications, hyperlipidemia, hypertension, and a history of renal dialysis. The most recent lab results show improved hemoglobin and hematocrit levels (9.6 g/dL and 30.0% respectively), which suggests that the anemia may be improving, possibly due to recent PRBC transfusions noted in the discharge note. However, the patient's history includes chronic iron deficiency anemia secondary to blood loss, which is a likely ongoing issue. The current admission is via emergency, and the patient has a history of gastrointestinal bleeding and endoscopic procedures, which could be related to the anemia. The most recent lab results show normal or near-normal values for most parameters, but the patient's history and the context of the admission suggest that the primary issue is likely related to the chronic anemia. The patient's other conditions are also significant, but the anemia is the most pressing issue given the recent transfusions and the context of the admission. Therefore, the projected diagnosis is likely to be iron deficiency anemia secondary to blood loss.

The primary diagnosis of this admission is D50.0-Iron deficiency anemia secondary to blood loss (chronic)

Multiple choice, single diagnosis (T3)

Input: Below is a history of a patient: Patient is a 20.0-year-old F.
Procedures performed: Esophagogastroduodenoscopy [EGD] with closed biopsy (ICD code: 4516); Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach (ICD code: 02HV33Z); Removal of Infusion Device from Great Vessel, External Approach

(ICD code: 02PYX3Z); Removal of Feeding Device from Stomach, External Approach (ICD code: 0DP6XUZ); Insertion of Infusion Device into Right Brachial Vein, Percutaneous Approach (ICD code: 05H933Z); Insertion of Infusion Device into Superior Vena Cava, Percutaneous Appr.
Admitted on 2188-06-01 07:20:00 via EU OBSERVATION from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2188-06-01 07:02:00 is: INDICATION: ___ female with DKA. TECHNIQUE: Chest PA and lateral COMPARISON: None available. FINDINGS: PA and lateral chest radiograph demonstrates clear lungs bilaterally with no focal opacity concerning for pneumonia. There is no pleural effusion or pneumothorax. Cardiomediastinal and hilar contours are within normal limits. No osseous abnormality is identified. IMPRESSION: No acute in thoracic abnormality. No foca.
All diagnoses from this admission: E10.65-Type 1 diabetes mellitus with hyperglycemia; Z79.4-Long term (current) use of insulin; E10.65-Type 1 diabetes mellitus with hyperglycemia; K31.84-Gastroparesis; E10.69-Type 1 diabetes mellitus with other specified complication; L98.8-Other specified disorders of the skin and subcutaneous tissue.

What is the projected diagnosis of the patient at this time?

A. E10.9-Type 1 diabetes mellitus without complications

B. E10.21-Type 1 diabetes mellitus with diabetic nephropathy

C. E10.11-Type 1 diabetes mellitus with ketoacidosis with coma

D. E10.10-Type 1 diabetes mellitus with ketoacidosis without coma

E. E10.8-Type 1 diabetes mellitus with unspecified complications

Include the answer in \boxed{} with a single letter. You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within tags. The final answer MUST BE put in \boxed{}.
Ground Truth: D
Model Output: I need to determine the projected diagnosis for this patient based on the provided history and current clinical context. The patient is a 20-year-old female with a history of Type 1 diabetes mellitus (E10.65) and has been admitted multiple times for DKA (Diabetic Ketoacidosis). The most recent admission was for DKA with worsening right leg pain, and there is concern for osteomyelitis. Looking at the lab results from the most recent admission (2190-05-09), the glucose is 335.0 mg/dL, which is elevated, and the anion gap is 16.0 mEq/L, which is within normal limits (normal anion gap is typically 8-16 mEq/L), but the patient has a history of DKA. The pH is 7.37, which is slightly acidic but not severely acidotic.
The question is asking for the projected diagnosis at this time. The patient has a history of DKA and is currently presenting with DKA and possible infection (cellulitis) and osteomyelitis. The ICD-10-CM codes provided in the history include E10.65 (Type 1 diabetes mellitus with hyperglycemia) and E10.69 (Type 1 diabetes mellitus with other specified complication), but the question is asking for the current projected diagnosis.
The options are: A. E10.9-Type 1 diabetes mellitus without complications B. E10.21-Type 1 diabetes mellitus with diabetic nephropathy C. E10.11-Type 1 diabetes mellitus with ketoacidosis with coma D. E10.10-Type 1 diabetes mellitus with ketoacidosis without coma E. E10.8-Type 1 diabetes mellitus with unspecified complications
The patient has DKA, but there is no mention of coma in the history or current presentation. The pH is 7.37, which is not indicative of coma. Therefore, the most appropriate code would be E10.10, which is Type 1 diabetes mellitus with ketoacidosis without coma.

D

Multiple choice, survival likelihood (T4)
Input: Below is a history of a patient: Patient is a 87.0-year-old F.
Procedures performed: Incision with removal of foreign body or device from skin and subcutaneous tissue (ICD code: 8605); Endoscopic sphincterotomy and papillotomy (ICD code: 5185); Other percutaneous procedures on biliary tract (ICD code: 5198); Other endoscopy of small intestine (ICD code: 4513); Diagnostic ultrasound of digestive system (ICD code: 8874); Endoscopic retrograde cholangiopancreatography [ERCP] (ICD code: 5110); Percutaneous biopsy of gallbladder or bile ducts (ICD code: 5112); Percutaneous hepatic ch.
Lab results: Basophils: 0.4 % (2194-07-28 07:15:00); Eosinophils: 2.4 % (2194-07-28 07:15:00); Hematocrit: 30.4 % (2194-08-04 04:18:00); Hemoglobin: 9.7 g/dL (2194-08-04 04:18:00); Lymphocytes: 10.2 % (2194-07-28 07:15:00); MCH: 29.1 pg (2194-08-04 04:18:00); MCHC: 31.9 % (2194-08-04 04:18:00); MCV: 91.0 fL (2194-08-04 04:18:00); Monocytes: 4.6 % (2194-07-28 07:15:00); Neutrophils: 82.4 % (2194-07-28 07:15:00); Platelet Count: 198.0 K/uL (2194-08-04 04:18:00); RDW: 14.6 % (2194-08-04 04:18:00); Red Blood Cel.
Admitted on 2194-07-27 09:26:00 via EW EMER. from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2194-07-29 17:13:00 is: ULTRASOUND DATED ___

INDICATION: choledocholithiasis, status post ERCP, now with renal failure, evaluate for obstruction.
COMPARISON: Comparison is made to previous CT dated ___.
FINDINGS: The right kidney measures 9.8 cm. The left kidney measures 11.1 cm. Both kidneys are normal in size and echogenicity with no evidence of hydronephrosis, suspicious renal lesion, or calculus identified. The urinary bladder is normal in .
Will the patient survive for at least 48 hours? Answer with "yes" or "no".
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within tags. The final answer MUST BE put in \boxed{}.
Ground Truth: yes
Model Output: I need to determine if the patient will survive for at least 48 hours. Looking at the patient's history, they are an 87-year-old female with a complex medical history including hypertension, diabetes, anemia, and a history of multiple procedures including ERCP and percutaneous interventions. They were admitted for abdominal pain and fever, with a diagnosis of ileus and later renal failure. The lab results show anemia (Hb 9.7 g/dL) and elevated neutrophils (82.4%), which could indicate infection or inflammation. The radiology notes show no free air and no evidence of obstruction, which is a positive sign. The patient has a central venous line and a device in the chest, which suggests they are being managed in a critical care setting. Given the patient's age and comorbidities, they are at high risk, but the absence of free air and the presence of a central line for support suggest that they are being actively managed. The patient's current condition is not immediately life-threatening, and with appropriate medical management, they are likely to survive for at least 48 hours. Therefore, the answer is "yes".

yes

Multiple choice, length of stay (T5)
Input: Below is a history of a patient: Patient is a 42.0-year-old M.
Procedures performed: Laparoscopic appendectomy (ICD code: 4701).
Admitted on 2202-07-05 20:07:00 via EU OBSERVATION from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2202-07-05 17:05:00 is: EXAM: Chest frontal and lateral views. CLINICAL INFORMATION: ___ male with history of cough. ___. FINDINGS: Frontal and lateral views of the chest were obtained. No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are stable and unremarkable. Mild multilevel degenerative changes of the visualized thoracic spine are seen. IMPRESSION: No acute cardiopulmon.
All diagnoses from this admission: R53.83-Other fatigue; E86.0-Dehydration; R11.2-Nausea with vomiting, unspecified; R19.7-Diarrhea, unspecified.
Patient is a 42.0-year-old M.
Procedures performed: Laparoscopic appendectomy (ICD code: 4701).
Lab results: Anion Gap: 15.0 mEq/L (2205-01-07 06:00:00); Bicarbonate: 23.0 mEq/L (2205-01-07 06:00:00); Chloride: 105.0 mEq/L (2205-01-07 06:00:00); Creatinine: 1.2 mg/dL (2205-01-07 06:00:00); Potassium: 4.4 mEq/L (2205-01-07 06:00:00); Sodium: 139.0 mEq/L (2205-01-07 06:00:00); Urea Nitrogen: 10.0 mg/dL (2205-01-07 06:00:00); Glucose: 85.0 mg/dL (2205-01-07 06:00:00); Hematocrit: 39.0 % (2205-01-07 06:00:00); Hemoglobin: 13.3 g/dL (2205-01-07 06:00:00); MCH: 30.6 pg (2205-01-07 06:00:00); MCHC: 34.1 % (22.
Admitted on 2205-01-05 14:50:00 via EW EMER. from EMERGENCY ROOM.
How long will the patient stay in the hospital?
A. 0-4 days
B. 5-8 days
C. 9-12 days
D. more than 12 days
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within tags. The final answer MUST BE put in \boxed{}.
Ground Truth: A
Model Output: I need to determine how long the patient will stay in the hospital based on the provided information. The patient is a 42-year-old male who was admitted on 2205-01-05 at 14:50 via emergency room. The admission note mentions he has a history of cough and was admitted for observation. The radiology notes show no acute cardiopulmonary issues, and the lab results are mostly within normal limits, with only mild dehydration and nausea/vomiting/diarrhea noted. The patient had a laparoscopic appendectomy previously, but that was not an issue during this admission. There is no indication of severe illness or complications that would require a prolonged hospital stay. The patient's symptoms (cough, fatigue, dehydration, nausea, diarrhea) are relatively mild and could be managed on an

outpatient basis or with a short inpatient stay. Given the lack of severe findings and the patient's age and general condition, a stay of 0-4 days is most likely.

A

G Hallucination Examples on MIMIC-IV Data

Hallucination Label: INTRINSIC, FABRICATED CLINICAL ENTITY
Input: Below is a history of a patient: Patient is a 91.0-year-old M.
Procedures performed: Extirpation of Matter from Common Bile Duct, Via Natural or Artificial Opening Endoscopic (ICD code: 0FC98ZZ).
Lab results: Alanine Aminotransferase (ALT): 21.0 IU/L (2187-07-22 05:45:00); Alkaline Phosphatase: 249.0 IU/L (2187-07-22 05:45:00); Anion Gap: 14.0 mEq/L (2187-07-22 05:45:00); Asparate Aminotransferase (AST): 15.0 IU/L (2187-07-22 05:45:00); Bicarbonate: 29.0 mEq/L (2187-07-22 05:45:00); Bilirubin, Total: 0.3 mg/dL (2187-07-22 05:45:00); Chloride: 105.0 mEq/L (2187-07-22 05:45:00); Creatinine: 1.2 mg/dL (2187-07-22 05:45:00); Glucose: 99.0 mg/dL (2187-07-22 05:45:00); Potassium: 4.9 mEq/L (2187-07-22 05:4. Admitted on 2187-07-19 15:58:00 via EW EMER. from EMERGENCY ROOM.
Discharge notes: The discharge note observed at chart time 2187-07-22 00:00:00 is: Name: ___ Unit No: ___ Admission Date: ____ Discharge Date: ___ Date of Birth: ____ Sex: M Service: MEDICINE Allergies: No Known Allergies / Adverse Drug Reactions Attending: ___. Chief Complaint: cough Major Surgical or Invasive Procedure: NONE History of Present Illness: Mr. ___ is a ___ year old male with PMH of dementia, stroke in ___, cholecystitis c/b CBD.
Radiology notes: The radiology note observed at chart time 2187-07-19 14:05:00 is: EXAMINATION: Chest: Frontal and lateral views INDICATION: History: ___ with cough // pna? TECHNIQUE: Chest: Frontal and Lateral COMPARISON: ___ at outside institution FINDINGS: Right base opacity is worrisome for pneumonia or aspiration. There is slight blunting of the right costophrenic angle which could be due to a trace pleural effusion. No pneumothorax is seen. The cardiac and mediastinal silhouettes are stable.
All diagnoses from this admission: J69.0-Pneumonitis due to inhalation of food and vomit; G93.41-Metabolic encephalopathy; E87.2-Acidosis; M31.6-Other giant cell arteritis; G30.9-Alzheimer's disease, unspecified; B37.2-Candidiasis of skin and nail; F01.50-Vascular dementia, unspecified severity, without behavioral disturbance, psychotic disturbance, mood disturbance, and anxiety; Z86.73-Personal history of transient ischemic attack (TIA), and cerebral infarction without residual deficits; Z85.118-Personal history of other malignant neoplasm of bronchus and lung; Z85.828-Personal history of other malignant neoplasm of skin; Z87.891-Personal history of nicotine dependence; Z78.1-Physical restraint status.
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission.
Ground Truth: All diagnoses from this admission: R53.1-Weakness; R53.81-Other malaise; R03.0-Elevated blood-pressure reading, without diagnosis of hypertension.
Model Output: I need to analyze the patient's history and current admission to determine the most likely diagnoses. The patient is a 91-year-old male with a history of dementia, stroke, and cholecystitis/c/b CBD (common bile duct). He was admitted for cough, and the radiology notes mention a right base opacity suspicious for pneumonia or aspiration, with a slight blunting of the right costophrenic angle. The discharge note lists several diagnoses, including pneumonitis due to inhalation of food and vomit, metabolic encephalopathy, acidosis, giant cell arteritis, Alzheimer's disease, candidiasis, vascular dementia, and personal history of various conditions.
Looking at the current admission, the patient's chief complaint is cough, and the radiology findings suggest a possible aspiration pneumonia. The lab results show normal liver function tests (ALT, AST, bilirubin) and normal creatinine, but elevated alkaline phosphatase, which could be related to his history of bile duct issues. The patient's history of dementia and stroke could contribute to aspiration risk.
The most likely diagnoses for this current admission, based on the chief complaint of cough and the radiology findings, would be pneumonia (likely aspiration pneumonia given the patient's history and the radiology note), metabolic encephalopathy (given the patient's history of dementia and the discharge note), and possibly acidosis (also listed in the discharge note).
All diagnoses from this admission: J69.0-Pneumonitis due to inhalation of food and vomit; G93.41-Metabolic encephalopathy; E87.2-Acidosis

The hallucination is identified as a fabricated clinical entity because the suggested diagnoses by the model does not align with the ground truth. This is an example of a case where the ground truth was focused on the primary reasons why the patient was admitted, but may not capture the main context for the patient's situation.

Hallucination Label: INTRINSIC, ATTRIBUTE MISALIGNMENT

Input: Below is a history of a patient: Patient is a 49.0-year-old F.
Procedures performed: Percutaneous abdominal drainage (ICD code: 5491); Venous catheterization, not elsewhere classified (ICD code: 3893).
Lab results: Basophils: 0.4 % (2187-08-18 06:35:00); Eosinophils: 3.7 % (2187-08-18 06:35:00); Hematocrit: 32.6 % (2187-08-18 06:35:00); Hemoglobin: 10.6 g/dL (2187-08-18 06:35:00); Lymphocytes: 30.5 % (2187-08-18 06:35:00); MCH: 32.0 pg (2187-08-18 06:35:00); MCHC: 32.4 % (2187-08-18 06:35:00); MCV: 99.0 fL (2187-08-18 06:35:00); Monocytes: 7.5 % (2187-08-18 06:35:00); Neutrophils: 58.0 % (2187-08-18 06:35:00); Platelet Count: 263.0 K/uL (2187-08-18 06:35:00); RDW: 12.5 % (2187-08-18 06:35:00); Red Blood Ce.
Admitted on 2187-08-16 20:00:00 via EW EMER. from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2187-08-17 10:07:00 is: CLINICAL INDICATION: Perforated diverticulitis and perisigmoid fluid collection. PROCEDURE: CT-guided drain placement. PHYSICIANS: ___. COMPLICATIONS: None immediate. PROCEDURE NOTE: Following a discussion of the risks, benefits and alternatives of the procedure, informed written consent was obtained. The patient was placed on the CT table in the supine position and a timeout procedure was performed utilizing three patie.
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission.
Ground Truth: All diagnoses from this admission: K57.32-Diverticulitis of large intestine without perforation or abscess without bleeding; K65.1-Peritoneal abscess; None.
Model Output: I need to analyze the patient's history to determine the most likely diagnoses for their current admission. The patient is a 49-year-old female admitted via emergency room on 2187-08-16. The radiology note from 2187-08-17 explicitly states the clinical indication was "Perforated diverticulitis and perisigmoid fluid collection," and a CT-guided drain was placed. This is a key finding.
The procedures performed include percutaneous abdominal drainage (ICD 5491) and venous catheterization (ICD 3893), which are consistent with managing an abdominal infection or abscess. The lab results show normal platelet count, but the hemoglobin is 10.6 g/dL (low-normal), hematocrit is 32.6% (low-normal), and MCV is 99.0 fL (slightly high), which could suggest mild anemia or early iron deficiency, but not necessarily acute or severe. The neutrophil percentage is elevated at 58%, which is consistent with an inflammatory or infectious process. The eosinophil percentage is 3.7%, which is within normal limits.
The patient's admission was via emergency room, and the radiology note indicates a perforated diverticulitis, which is a common cause of acute abdomen in older adults. The perisigmoid fluid collection suggests an abscess or localized infection around the sigmoid colon. The CT-guided drain placement is a standard intervention for such cases.
Other possible diagnoses could include appendicitis, but the radiology note specifically points to diverticulitis. There is no mention of other symptoms or findings that would suggest other conditions. The lab results do not show signs of severe infection (e.g., very high WBC or CRP, though CRP is not provided). The patient's age and gender are consistent with diverticulitis being a common condition.
Therefore, the most likely diagnosis is perforated diverticulitis with abscess formation. Other possible diagnoses could include secondary peritonitis, but that is a complication of the diverticulitis. The patient's lab results do not suggest other systemic infections or conditions.
All diagnoses from this admission: Perforated diverticulitis, Perisigmoid abscess, Secondary peritonitis

The hallucination occurs where the model labels the admission diagnosis as perforated diverticulitis and secondary peritonitis, misaligning disease severity and complications relative to the ground truth, which specifies non-perforated diverticulitis with peritoneal abscess only.

Hallucination Label: EXTRINSIC, SEMANTIC-PHYSIOLOGY MISMATCH

Input: Below is a history of a patient: Patient is a 65.0-year-old M.
Procedures performed: Closed [endoscopic] biopsy of bronchus (ICD code: 3324); Biopsy of lymphatic structure (ICD code: 4011); Injection or infusion of cancer chemotherapeutic substance (ICD code: 9925).
Lab results: Anion Gap: 16.0 mEq/L (2191-12-17 06:00:00); Bicarbonate: 25.0 mEq/L (2191-12-17 06:00:00); Calcium, Total: 9.0 mg/dL (2191-12-17 06:00:00); Chloride: 103.0 mEq/L (2191-12-17 06:00:00); Creatinine: 0.7 mg/dL (2191-12-17 06:00:00); Potassium: 4.3 mEq/L (2191-12-17 06:00:00); Sodium: 140.0 mEq/L (2191-12-17 06:00:00); Urea Nitrogen: 16.0 mg/dL (2191-12-17 06:00:00); Glucose: 93.0 mg/dL (2191-12-17 06:00:00); Hematocrit: 39.2 % (2191-12-16 07:20:00); Hemoglobin: 13.2 g/dL (2191-12-16 07:20:00); MCH.
Admitted on 2191-12-14 22:10:00 via EW EMER. from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2191-12-15 17:05:00 is: HISTORY: ___ year old man with newly diagnosed Pancoast tumor; evaluate for metastatic disease. TECHNIQUE: Multiplanar multi sequence pre- and post-contrast MR images of the brain were obtained. COMPARISON: None. FINDINGS: There is no evidence of acute intracranial infarct or hemorrhage. There are

nonspecific small scattered T2/FLAIR high signal foci throughout the brain consistent which may be sequela of chronic microangi.
Discharge notes: The discharge note observed at chart time 2191-12-17 00:00:00 is: Name: ___ Unit No: ___ Admission Date: ___ Discharge Date: ___ Date of Birth: ____ Sex: M Service: MEDICINE Allergies: No Known Allergies / Adverse Drug Reactions Attending: ___ Chief Complaint: Difficulty walking and left sided chest/arm pain Major Surgical or Invasive Procedure: Flexible bronchoscopy, EBUS with biopsy ___ History of Present Illness: ___. There were no radiology notes taken at this admission.
All diagnoses from this admission: D49.1-Neoplasm of unspecified behavior of respiratory system; J44.9-Chronic obstructive pulmonary disease, unspecified; Z87.891-Personal history of nicotine dependence; F10.10-Alcohol abuse, uncomplicated; K21.9-Gastro-esophageal reflux disease without esophagitis; E03.9-Hypothyroidism, unspecified; F32.9-Major depressive disorder, single episode, unspecified.
[Summary of the sequential encounter.]
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission.
Ground Truth: All diagnoses from this admission: A41.9-Sepsis, unspecified organism; J18.9-Pneumonia, unspecified organism; G92.8-Other toxic encephalopathy; J44.0-Chronic obstructive pulmonary disease with (acute) lower respiratory infection; C34.32-Malignant neoplasm of lower lobe, left bronchus or lung; J96.11-Chronic respiratory failure with hypoxia; I27.82-Chronic pulmonary embolism; L97.229-Non-pressure chronic ulcer of left calf with unspecified severity; Z86.16-Personal history of COVID-19; I10-Essential (primary) hypertension; E78.5-Hyperlipidemia, unspecified; K21.9-Gastro-esophageal reflux disease without esophagitis; E03.9-Hypothyroidism, unspecified; G60.8-Other hereditary and idiopathic neuropathies; Z87.891-Personal history of nicotine dependence; E55.9-Vitamin D deficiency, unspecified; Z51.5-Encounter for palliative care; B35.3-Tinea pedis; Z99.81-Dependence on supplemental oxygen; I87.8-Other specified disorders of veins.
Model Output: I need to analyze the patient's history and current admission to determine the most likely diagnoses. The patient is a 65-year-old male with a history of cancer (Pancoast tumor), chronic obstructive pulmonary disease (COPD), and recent chemotherapy. The current admission is for difficulty walking and left-sided chest/arm pain. The X-ray shows a large mass in the left lung, which is consistent with the known Pancoast tumor. The patient also has a history of pulmonary embolism and pneumonia.
The lab results show normal electrolytes and creatinine, but the pH is 5.5, which is very low and indicates severe acidosis. This could be due to the cancer or the chemotherapy. The most likely diagnoses for this patient are the Pancoast tumor, COPD, and the acidosis. The other diagnoses are less likely because they are not directly related to the current admission.
All diagnoses from this admission: D49.1-Neoplasm of unspecified behavior of respiratory system; J44.9-Chronic obstructive pulmonary disease, unspecified; Z87.891-Personal history of nicotine dependence; F10.10-Alcohol abuse, uncomplicated; K21.9-Gastro-esophageal reflux disease without esophagitis; E03.9-Hypothyroidism, unspecified; F32.9-Major depressive disorder, single episode, unspecified

The semantic-physiology hallucination here is due to the misunderstanding of the hallucination categorizer. The chest-x-ray is not given as input to the categorizer, so it falsely identifies a hallucination where the model output references a non-existent entity.

Hallucination Label: INTRINSIC, CROSS-PATIENT OR TEMPLATE LEAKAGE
Input: Below is a history of a patient: Patient is a 66.0-year-old F.
Procedures performed: Supplement Abdominal Wall with Synthetic Substitute, Open Approach (ICD code: 0WUF0JZ); Release Right Abdomen Muscle, Open Approach (ICD code: 0KNK0ZZ); Release Left Abdomen Muscle, Open Approach (ICD code: 0KNL0ZZ); Excision of Abdomen Skin, External Approach (ICD code: 0HB7XZZ); Removal of Tissue Expander from Trunk Subcutaneous Tissue and Fascia, Open Approach (ICD code: 0JPT0NZ).
Lab results: Anion Gap: 15.0 mEq/L (2187-03-20 23:15:00); Bicarbonate: 27.0 mEq/L (2187-03-20 23:15:00); Chloride: 106.0 mEq/L (2187-03-20 23:15:00); Creatinine: 0.6 mg/dL (2187-03-20 23:15:00); Glucose: 131.0 mg/dL (2187-03-20 23:15:00); Potassium: 4.6 mEq/L (2187-03-20 23:15:00); Sodium: 143.0 mEq/L (2187-03-20 23:15:00); Urea Nitrogen: 24.0 mg/dL (2187-03-20 23:15:00); Atypical Lymphocytes: 6.0 % (2187-03-20 23:15:00); Bands: 1.0 % (2187-03-20 23:15:00); Basophils: 0.0 % (2187-03-20 23:15:00); Eosinophils.
Admitted on 2187-03-21 01:04:00 via EU OBSERVATION from EMERGENCY ROOM.
Radiology notes: The radiology note observed at chart time 2187-03-21 00:09:00 is: INDICATION: ___ woman with chest pain. COMPARISON: ___. FINDINGS: Cardiac, mediastinal and hilar contours are within normal limits. Both lungs are clear, with no focal consolidation, pleural effusion or pneumothorax. A tube is noted extending from the chest region into the abdominal region on the lateral view and is likely external to the patient. IMPRESSION: No acute cardiopulmonary process.
[Summary of the sequential encounter.]
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission.

Ground Truth: All diagnoses from this admission: J18.9-Pneumonia, unspecified organism; G92-Toxic encephalopathy; N17.9-Acute kidney failure, unspecified; E87.0-Hyperosmolality and hypernatremia; F60.7-Dependent personality disorder; F31.9-Bipolar disorder, unspecified; G40.909-Epilepsy, unspecified, not intractable, without status epilepticus; M65.839-Other synovitis and tenosynovitis, unspecified forearm; M25.373-Other instability, unspecified ankle; Z23-Encounter for immunization; Z87.81-Personal history of (healed) traumatic fracture; Q68.8-Other specified congenital musculoskeletal deformities; I10-Essential (primary) hypertension; E78.5-Hyperlipidemia, unspecified; E11.9-Type 2 diabetes mellitus without complications; Z79.84-Long term (current) use of oral hypoglycemic drugs; M25.519-Pain in unspecified shoulder; G89.29-Other chronic pain; Z91.14-Patient's other noncompliance with medication regimen; None; Y92.039-Unspecified place in apartment as the place of occurrence of the external cause; R41.0-Disorientation, unspecified; K46.9-Unspecified abdominal hernia without obstruction or gangrene.
Model Output: I need to analyze the patient's history, lab results, and radiology reports to determine the most likely diagnoses for their current admission. The patient is a 66-year-old female with a history of abdominal surgeries and current symptoms of chest pain and possible pneumonia. The chest X-ray shows improved left lower lobe pneumonia and vascular congestion , which suggests a respiratory infection that is resolving. The patient also has a history of hypertension and hypercholesterolemia, which are risk factors for cardiovascular disease. The lab results show normal electrolytes and renal function, with a slightly elevated anion gap and normal glucose levels. The patient's current admission is for evaluation of possible pneumonia and pulmonary edema, with a history of chest pain. The most likely diagnoses for this patient are pneumonia, pulmonary edema, and hypertension. The patient's history of abdominal surgeries and the presence of a tube extending from the chest to the abdomen suggest a possible complication from previous surgery, but the current X-ray shows no acute cardiopulmonary process. The patient's current symptoms and lab results are consistent with a resolving pneumonia and pulmonary edema. The patient's history of hypertension and hypercholesterolemia increases the risk of cardiovascular disease, but the current X-ray shows no evidence of acute cardiopulmonary process. Therefore, the most likely diagnoses for this patient are pneumonia, pulmonary edema, and hypertension.
All diagnoses from this admission: R07.9-Chest pain, unspecified; I10-Essential (primary) hypertension; E78.0-Pure hypercholesterolemia

The hallucination is identified as cross-patient template leakage. To note, the chest x-rays seen by QoQ-Med are not seen by the hallucination categorizer, gpt-5-mini. The radiology note states that "both lungs are clear", but the model output states that "the chest x-ray shows improved..." This was likely flagged by gpt-5-mini due to the fact that the model's claim seems unfounded. Moreover, words like "consistent" can be flagged due to seeming as though the model is following memorized patterns, instead of being specific to the patient's case.

Hallucination Label: INTRINSIC, NEGATION OR ABSENCE FAILURE
Input: Below is a history of a patient: Patient is a 52.0-year-old F.
Procedures performed: Performance of Urinary Filtration, Intermittent, Less than 6 Hours Per Day (ICD code: 5A1D70Z); Drainage of Left Hip Joint, Open Approach (ICD code: 0S9B0ZZ); Drainage of Left Hip Bursa and Ligament, Open Approach (ICD code: 0M9M0ZZ); Inspection of Lower Intestinal Tract, Via Natural or Artificial Opening Endoscopic (ICD code: 0DJD8ZZ); Insertion of Tunneled Vascular Access Device into Chest Subcutaneous Tissue and Fascia, Percutaneous Approach (ICD code: 0JH63XZ); Insertion of Infusion Device i.
Lab results: Anion Gap: 16.0 mEq/L (2189-10-28 07:30:00); Bicarbonate: 29.0 mEq/L (2189-10-28 07:30:00); Bilirubin, Total: 0.2 mg/dL (2189-10-21 06:55:00); Calcium, Total: 9.6 mg/dL (2189-10-28 07:30:00); Chloride: 97.0 mEq/L (2189-10-28 07:30:00); Creatine Kinase (CK): 22.0 IU/L (2189-10-16 06:35:00); Creatinine: 6.0 mg/dL (2189-10-28 07:30:00); Glucose: 89.0 mg/dL (2189-10-28 07:30:00); Lactate Dehydrogenase (LD): 233.0 IU/L (2189-10-21 06:55:00); Magnesium: 2.1 mg/dL (2189-10-28 07:30:00); Phosphate: 3.6.
Admitted on 2189-10-14 18:20:00 via EW EMER. from PROCEDURE SITE.
Radiology notes: The radiology note observed at chart time 2189-10-15 15:10:00 is: INDICATION: ___ woman with recent dialysis line placement and mediastinal widening. Comparison is made to the prior study of ___. PA AND LATERAL RADIOGRAPH OF THE CHEST: The heart size and hilar contours are normal. There has been mild interval reduction in the mediastinal widening. Lungs are clear with no focal consolidation, pleural effusion or pneumothorax. The hemodialysis catheter line is unchanged in location. IMPRESSIO.
What is the projected diagnosis of the patient at this time?
A. N99.83-Residual ovary syndrome
B. N99.821-Postprocedural hemorrhage of a genitourinary system organ or structure following other procedure
C. N99.820-Postprocedural hemorrhage of a genitourinary system organ or structure following a genitourinary system procedure
D. N99.81-Other intraoperative complications of genitourinary system
E. N99.85-Post endometrial ablation syndrome

Ground Truth: B
Model Output: I need to determine the projected diagnosis for this 52-year-old female patient based on the provided history, lab results, procedures, and the radiology note.
The patient has a history of end-stage renal disease (N18.6), hypertension (I10), systemic lupus erythematosus (M32.10), and hyperlipidemia (E78.5). She has been on intermittent hemodialysis and has had multiple procedures, including the insertion of a tunneled vascular access device into the chest subcutaneous tissue and fascia (0JH63XZ), which is likely a hemodialysis catheter.
The radiology note from October 15, 2189, states that the patient is a woman with recent dialysis line placement and mediastinal widening. The comparison is made to a prior study, and the findings include normal heart size and hilar contours, clear lungs, and no focal consolidation, pleural effusion, or pneumothorax. The hemodialysis catheter line is unchanged in location. The impression is not explicitly stated, but the context suggests that the mediastinal widening is being monitored.
The question asks for the projected diagnosis at this time, and the options are all related to postprocedural hemorrhage or complications of the genitourinary system. However, the patient's history and the radiology note do not mention any genitourinary system procedures or complications.
The procedures listed are related to the hip, lower intestinal tract, and vascular access device insertion.
The radiology note mentions "recent dialysis line placement" and "mediastinal widening," which is likely related to the hemodialysis catheter. The mediastinal widening could be due to the catheter or a complication such as a hematoma, but the note states that there has been a "mild interval reduction in the mediastinal widening," suggesting that the condition is improving.
None of the options (A, B, C, D, E) are directly related to the patient's history or the radiology findings. Option A (N99.83-Residual ovary syndrome) is not relevant to the patient's history. Option B (N99.821-Postprocedural hemorrhage of a genitourinary system organ or structure following other procedure) and Option C (N99.820-Postprocedural hemorrhage of a genitourinary system organ or structure following a genitourinary system procedure) are related to genitourinary system hemorrhage, but the patient's history does not indicate any genitourinary system procedures. Option D (N99.81-Other intraoperative complications of genitourinary system) is also not relevant. Option E (N99.85-Post endometrial ablation syndrome) is not relevant either.
Given the lack of direct relevance of the options to the patient's history and the radiology findings, I need to consider whether there is any indirect connection. The patient has systemic lupus erythematosus, which can cause a variety of complications, but none of the options are related to lupus.
The most likely answer is that there is no direct connection, but the question asks for the projected diagnosis, and the options are all related to genitourinary system complications. Therefore, I need to choose the option that is most likely to be related to the patient's condition.

The hallucination is identified as negation or absence failure. The model fails to recognize that even though the radiology note may not suggest genitourinary system procedures, that diseases relating to the genitourinary system should not be ruled out as a possible answer. It incorrectly treats the absence of documentation as evidence of absence of disease.

Hallucination Label: EXTRINSIC, CAUSAL ATTRIBUTION
Input: Below is a history of a patient: Patient is a 75.0-year-old M.
Procedures performed: Continuous invasive mechanical ventilation for less than 96 consecutive hours (ICD code: 9671); Insertion of intercostal catheter for drainage (ICD code: 3404); Cardiopulmonary resuscitation, not otherwise specified (ICD code: 9960); Central venous catheter placement with guidance (ICD code: 3897); Closed [endoscopic] biopsy of bronchus (ICD code: 3324); Hemodialysis (ICD code: 3995); Continuous invasive mechanical ventilation for less than 96 consecutive hours (ICD code: 9671); Hemodialysis (IC.
Lab results: Bicarbonate: 19.0 mEq/L (2185-06-15 06:20:00); Creatinine: 3.6 mg/dL (2185-06-15 06:20:00); Glucose: 85.0 mg/dL (2185-06-15 06:20:00); Potassium: 5.0 mEq/L (2185-06-15 06:20:00); Sodium: 136.0 mEq/L (2185-06-15 06:20:00); Urea Nitrogen: 56.0 mg/dL (2185-06-15 06:20:00); Hematocrit: 31.8 % (2185-06-15 06:20:00); Hemoglobin: 10.8 g/dL (2185-06-15 06:20:00); MCH: 30.2 pg (2185-06-15 06:20:00); MCHC: 33.8 % (2185-06-15 06:20:00); MCV: 89.0 fL (2185-06-15 06:20:00); Platelet Count: 284.0 K/uL (2185-0.
Admitted on 2185-06-13 18:02:00 via EW EMER. from EMERGENCY ROOM.
Discharge notes: The discharge note observed at chart time 2185-06-15 00:00:00 is: Name: ___ Unit No: ___ Admission Date: ___ Discharge Date: ___ Date of Birth: ____ Sex: M Service: MEDICINE Allergies: Bactrim Ds / Levaquin / Sulfa (Sulfonamides) Attending: ___. Chief Complaint: lightheadedness Major Surgical or Invasive Procedure: none History of Present Illness: Mr. ___ is a ___ year old male with history of hypertension and chronic ki.
Radiology notes: The radiology note observed at chart time 2185-06-13 17:38:00 is: HISTORY: ___ male with lightheadedness, to assess for a cardiopulmonary process. TECHNIQUE: AP and lateral radiographs of the chest were performed. Comparison is made with multiple prior radiographs dating back to ___. FINDINGS: There are multiple right-sided diaphragmatic eventrations. The lungs are

over-inflated suggestive of COAD. The contour at the right cardiophrenic angle is consistent with fat-filled Bochdalek's her.
All diagnoses from this admission: R55-Syncope and collapse; N18.4-Chronic kidney disease, stage 4 (severe); E86.0-Dehydration; I12.9-Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease; D63.1-Anemia in chronic kidney disease; K21.9-Gastro-esophageal reflux disease without esophagitis; D50.9-Iron deficiency anemia, unspecified; R05-Cough; Z85.048-Personal history of other malignant neoplasm of rectum, rectosigmoid junction, and anus; Z87.440-Personal history of urinary (tract) infections; Z85.51-Personal history of malignant neoplasm of bladder; None.
[Summary of the sequential encounter.]
What is the projected diagnosis of the patient at this time?
Ground Truth: The primary diagnosis of this admission is N17.9-Acute kidney failure, unspecified. The primary diagnosis of this admission is categorized as ICD Chapter 14: Diseases of the genitourinary system (N00-N99). To be more specific, it is within the Acute kidney failure and chronic kidney disease (N17-N19) block.
Model Output: I need to determine the projected diagnosis for this patient based on the provided history, lab results, and imaging.
Looking at the patient's history, they are a 75-year-old male with a history of hypertension and chronic kidney disease (CKD), as indicated in the discharge notes and lab results. The lab results show elevated creatinine (3.6 mg/dL and later 4.3 mg/dL) and urea nitrogen (56.0 mg/dL and later 31.0 mg/dL), which are consistent with renal impairment. The hematocrit and hemoglobin are low (31.8% and 10.8 g/dL, then 26.2% and 8.4 g/dL), which is consistent with anemia, likely due to CKD (as noted in the diagnosis D63.1-Anemia in chronic kidney disease).
The radiology notes from the first admission mention "multiple right-sided diaphragmatic eventrations" and "lungs are over-inflated suggestive of COAD" (Chronic Obstructive Airway Disease). However, the most recent radiology note (from 2187-05-25) is about an AV fistulogram and angioplasty, which is related to the patient's end-stage renal disease and hemodialysis needs.
The patient is admitted via emergency room with the chief complaint of lightheadedness.
The most recent admission (2187-05-20) is for a patient with end-stage renal disease who failed a hemodialysis session.
The radiology note for this admission is specifically about the AV fistula and angioplasty.
Given the patient's history of end-stage renal disease, the recent admission for failed hemodialysis, and the radiology note focusing on the AV fistula, the primary issue at this time is likely related to the vascular access for hemodialysis. The patient's lightheadedness could be due to hypotension from the failed dialysis session or other complications.
The projected diagnosis should reflect the most immediate and relevant condition based on the current admission. The patient is being evaluated for the AV fistula, and the primary concern is likely the failure of the hemodialysis session, which is a complication of end-stage renal disease.
Therefore, the primary diagnosis of this admission is likely related to the end-stage renal disease and the complications thereof, such as the failed hemodialysis session.
The primary diagnosis of this admission is End-stage renal disease with failed hemodialysis session

The model hallucinates an unsupported causal mechanism by attributing the patient's admission to a failed hemodialysis session and end-stage renal disease, despite no evidence of dialysis failure and an explicitly stated primary diagnosis of acute kidney failure.

Hallucination Label: EXTRINSIC, INCORRECT PATHOPHYSIOLOGICAL REASONING
Input: Below is a history of a patient: Patient is a 88.0-year-old M.
Procedures performed: Implantation of cardiac resynchronization pacemaker without mention of defibrillation, total system [CRT-P] (ICD code: 0050); Percutaneous transluminal coronary angioplasty [PTCA] (ICD code: 0066); Insertion of drug-eluting coronary artery stent(s) (ICD code: 3607); Left heart cardiac catheterization (ICD code: 3722); Coronary arteriography using two catheters (ICD code: 8856); Insertion of two vascular stents (ICD code: 0046); Procedure on single vessel (ICD code: 0040); Other and unspecified c.
Lab results: Creatinine: 1.4 mg/dL (2188-12-03 05:38:00); Urea Nitrogen: 23.0 mg/dL (2188-12-03 05:38:00); Hematocrit: 27.9 % (2188-12-03 05:38:00); Hemoglobin: 9.9 g/dL (2188-12-03 05:38:00); MCH: 34.9 pg (2188-12-03 05:38:00); MCHC: 35.6 % (2188-12-03 05:38:00); MCV: 98.0 fL (2188-12-03 05:38:00); Platelet Count: 139.0 K/uL (2188-12-03 05:38:00); RDW: 14.8 % (2188-12-03 05:38:00); Red Blood Cells: 2.84 m/uL (2188-12-03 05:38:00); White Blood Cells: 4.3 K/uL (2188-12-03 05:38:00); Anion Gap: 9.0 mEq/L (2188.
Admitted on 2188-12-02 19:37:00 via EW EMER. from PROCEDURE SITE.
Provide a numbered list of 3 to 10 most likely diagnoses for this patient for their current admission. The answer should start with "All diagnoses from this admission: "
Ground Truth: All diagnoses from this admission: I25.10-Atherosclerotic heart disease of native coronary artery without angina pectoris; I50.22-Chronic systolic (congestive) heart failure; I25.810-Atherosclerosis of coronary artery bypass graft(s) without angina

pectoris; I25.82-Chronic total occlusion of coronary artery; I50.9-Heart failure, unspecified; I35.9-Nonrheumatic aortic valve disorder, unspecified; M10.9-Gout, unspecified; N40.0-Benign prostatic hyperplasia without lower urinary tract symptoms; I25.2-Old myocardial infarction; Z98.61-Coronary angioplasty status.

Model Output: I need to analyze the patient's history to determine the most likely diagnoses for their current admission. The patient is an 88-year-old male with a history of cardiac procedures including implantation of a CRT-P pacemaker, PTCA, stent insertion, and cardiac catheterization. These suggest a history of significant cardiovascular disease, likely including coronary artery disease and heart failure.

The lab results show elevated creatinine (1.4 mg/dL) and urea nitrogen (23.0 mg/dL), which may indicate renal impairment. The hematocrit and hemoglobin are low (27.9% and 9.9 g/dL), suggesting anemia, which could be due to chronic disease, blood loss, or bone marrow suppression. The MCV is elevated (98.0 fL), indicating macrocytic anemia, which could be due to B12 or folate deficiency, or chronic disease. The RDW is also elevated (14.8%), which is consistent with anemia of chronic disease.

The patient was admitted via emergency from the procedure site, which suggests that the current admission is likely related to a complication or issue arising from the recent procedures. The most likely diagnoses would be related to the cardiovascular system, given the history of heart disease and procedures.

The most likely diagnoses for this patient's current admission are:

Heart failure with reduced ejection fraction (HFrEF) - given the history of CRT-P implantation and coronary artery disease.

Anemia of chronic disease - given the lab results showing low hemoglobin and hematocrit with elevated MCV and RDW.

Renal impairment - given the elevated creatinine and urea nitrogen levels.

These diagnoses are the most likely based on the patient's history and lab results. The patient's age and history of cardiovascular disease also increase the likelihood of these diagnoses.

All diagnoses from this admission: Heart failure with reduced ejection fraction, Anemia of chronic disease, Renal impairment

The model commits an extrinsic hallucination by drawing incorrect pathophysiological conclusions. For example, it equates mild lab value elevation with the clinically meaningful renal impairment, without further evidence of chronic kidney disease diagnosis.

References

1. Christiano, P. F. *et al.* Deep reinforcement learning from human preferences. *Adv. neural information processing systems* **30** (2017).

2. Ziegler, D. M. *et al.* Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

3. Naik, A., Wan, Y., Tomar, M. & Sutton, R. S. Reward centering. *arXiv preprint arXiv:2405.09999* (2024).