RADDIFF: AUTOMATED RADIOLOGY DISCOVERY FROM COHORT DIFFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many clinical questions in radiology are questions about cohorts: Which radiographic findings distinguish pneumonia patients who die from those who survive? How does COVID-19 present in older versus younger patients? What separates the radiographs a race classifier assigns to different groups? Answering them requires discovering, in open-ended language, what distinguishes two cohorts of hundreds of studies, a task that is laborious for radiologists and on which general-domain methods fail. We introduce RADDIFF, a multimodal agent for automated radiology discovery from cohort differences. RADDIFF runs a hypothesize–verify–refine loop: a proposer reads domain-adapted captions and images of a few sampled studies and proposes candidate differences, a ranker verifies every candidate on the full cohorts with a medical vision–language model, and the best-verified hypotheses, with zoomed-in views of the regions they refer to, are fed back to refine the next round. On RADDIFFBENCH, a new benchmark of 57 radiologist-validated cohort pairs from MIMIC-CXR, RADDIFF achieves 47.4% top-1 accuracy versus 1.8% for the general-domain VisDiff, with every component contributing. Applied to the three questions above, it surfaces radiologist-reviewed discoveries, from the device burden of pneumonia non-survivors to the non-anatomical cues behind a race classifier's predictions.

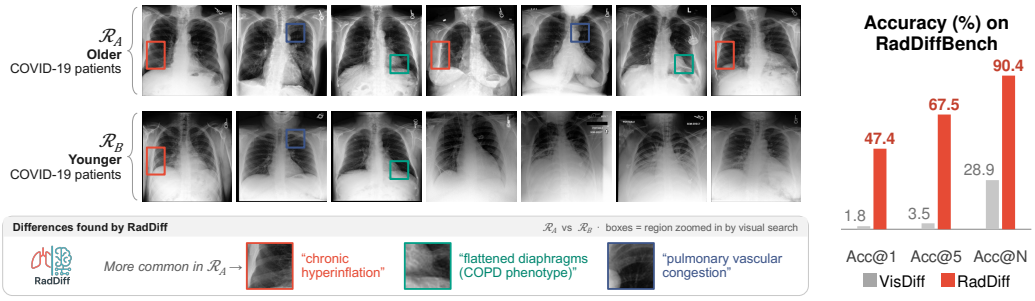


Figure 1: RADDIFF discovers what distinguishes two radiology cohorts. *Left:* for older ($\mathcal{R}_A$) vs. younger ($\mathcal{R}_B$) COVID-19 patients, RADDIFF returns ranked natural-language differences, each grounded in the regions it zoomed into during visual search (boxes). *Right:* accuracy on RADDIFFBENCH compared with VisDiff (Dunlap et al., 2024).

1 INTRODUCTION

Many consequential questions in radiology compare cohorts. Q1 Which radiographic patterns distinguish pneumonia patients who die in hospital from those who survive (Kim et al., 2023)? Q2 How does COVID-19 present differently with age (Okoye et al., 2022)? Q3 Deep networks predict patient race from chest radiographs with high accuracy (Gichoya et al., 2022); what do they rely on (Lotter, 2024)? Answers to such questions generate clinical hypotheses and expose the shortcuts of medical AI models, yet obtaining them today requires experts to read hundreds of studies per cohort and articulate what differs.

Existing tools do not support open-ended, cohort-level discovery. Comparing labels extracted from reports can only confirm differences anticipated in the label set. VisDiff (Dunlap et al., 2024)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

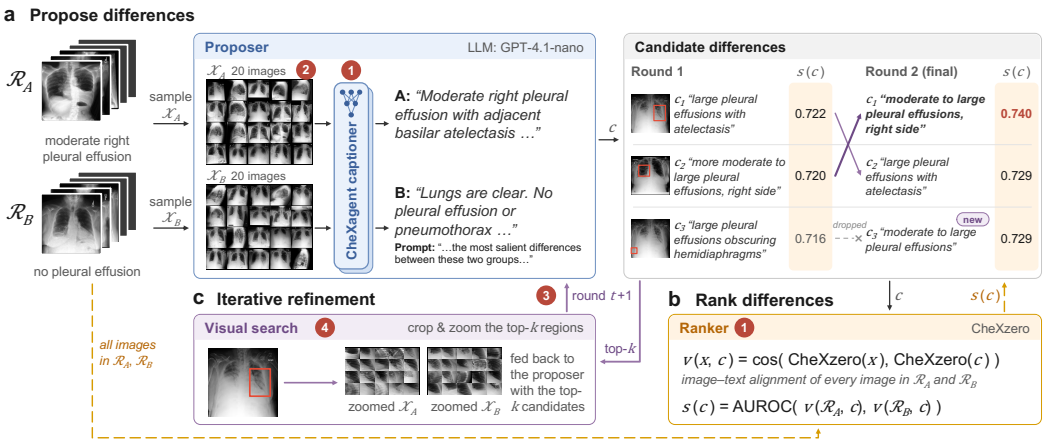


Figure 2: Overview of RADDIFF. (a) The proposer reads CheXagent captions and image grids of 20 sampled studies per cohort and proposes candidate differences. **(b)** The ranker verifies every candidate on the full cohorts by the AUROC of its CheXzero image–text scores. **(c)** The top-k verified hypotheses, with their scores and zoomed crops of the regions they refer to, are fed back to the proposer. Badges mark the four components: ❶ knowledge injection, ❷ multimodal reasoning, ❸ iterative refinement, and ❹ visual search.

describes how two image sets differ in natural language, proposing candidates from image captions with an LLM and ranking them with CLIP (Radford et al., 2021), but it was built for natural images and reaches only 1.8% top-1 accuracy on radiographs (Table 2). Radiographs defeat it in four ways: naming their findings requires domain knowledge; findings rarely survive compression into a generic caption; a single pass of reasoning cannot revise hypotheses the way radiologists do; and the differences that matter are often subtle and local.

RADDIFF automates discovery by emulating how radiologists compare cohorts (Figure 2): read a sample of cases, form hypotheses, check them against the whole population, and look again where it matters. A **proposer** reads small samples and proposes candidate differences, a **ranker** verifies each on the full cohorts, and the best-verified hypotheses are fed back to **refine** the next round. Four components address these gaps in turn: ❶ **knowledge injection** through the chest-radiograph models CheXagent (Chen et al., 2024) and CheXzero (Tiu et al., 2022); ❷ **multimodal reasoning** over the images alongside their captions; ❸ **iterative refinement** conditioned on verified hypotheses; and ❹ **visual search**, which zooms into the regions that support the leading hypotheses.

RADDIFF recovers known differences and surfaces new, clinically coherent ones. On RADDIFFBENCH, our benchmark of 57 MIMIC-CXR cohort pairs (Johnson et al., 2019) with radiologist-validated differences (e.g., *moderate right pleural effusion* vs. *no pleural effusion*), RADDIFF raises top-1 accuracy from 1.8% to 47.4%, and every component contributes. Applied to Q1–Q3, it finds that pneumonia non-survivors carry more devices and more extensive bilateral disease Q1, that older COVID-19 patients show chronic structural change whereas younger ones show acute opacities Q2, and that a race classifier's most confident predictions rely on artifacts like devices, acquisition context, and reported normality rather than anatomy Q3. Radiologists reviewed these findings and judged them clinically coherent. Our contributions are:

- **Task and benchmark:** we formulate automated radiology discovery as open-ended, cohort-level difference description and introduce the radiologist-validated RADDIFFBENCH (Section 3).
- **Method:** RADDIFF, a hypothesize–verify–refine agent that improves top-1 accuracy by 45.6 points over VisDiff (Sections 4–5).
- **Discoveries:** three radiologist-reviewed case studies on outcome prediction, age-dependent phenotypes, and the auditing of a medical AI model (Section 6).

2 RELATED WORK

Vision–language models and agents. Contrastive vision–language models (VLMs) such as CLIP align images and text (Radford et al., 2021; Girdhar et al., 2023; Zhai et al., 2023), and generative

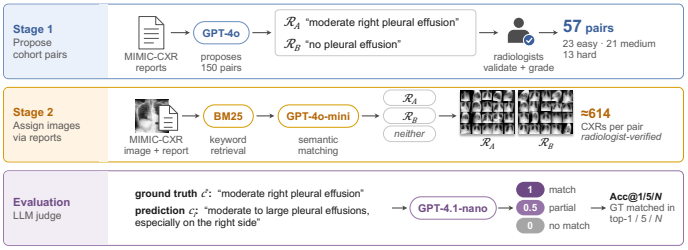


Figure 3: RADDIFFBENCH construction and evaluation. Stage 1: an LLM proposes difference concepts from reports and radiologists validate them. Stage 2: reports assign studies to cohorts. An LLM judge scores predictions against the ground truth.

Table 1: Key statistics.

Statistic	Number
Candidate pairs (GPT-4o)	150
Radiologist-validated pairs	**57**
Easy / Med. / Hard	23 / 21 / 13
Images per pair (mean)	**614**
Easy / Med. / Hard	614 / 607 / 625
range	418–684
Images per cohort (mean)	307
Image source	MIMIC-CXR

multimodal LLMs (MLLMs) reason over visual inputs (OpenAI, 2023; Liu et al., 2023; Li et al., 2023). Composed into agents, they solve more complex tasks (Wang et al., 2024; Shen et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023); VisDiff (Dunlap et al., 2024) combines CLIP and MLLMs to describe how two image sets differ. RADDIFF builds on its proposer–ranker structure and adapts it to radiology through domain knowledge, multimodal reasoning, iterative refinement, and visual search (Wu & Xie, 2024).

VLMs for radiology. Embedding models such as CheXzero (Tiu et al., 2022), BioViL (Boecking et al., 2022), GLoRIA (Huang et al., 2021), ConVIRT (Zhang et al., 2022), and MGCA (Wang et al., 2022), and generative models such as CheXagent (Chen et al., 2024), MAIRA-2 (Bannur et al., 2024), and MedGemma (Sellergren et al., 2025), perform well on visual question answering, disease classification, longitudinal analysis, and report generation, but reason about one study or patient at a time. We instead study *cohort-level* reasoning: discovering, in natural language, what distinguishes two populations of radiographs.

3 TASK AND BENCHMARK

3.1 TASK: COHORT DIFFERENCE DISCOVERY

Given two cohorts of radiographs, $\mathcal{R}_A$ and $\mathcal{R}_B$, the goal is to describe in natural language what distinguishes them. Following the set-difference formulation of Dunlap et al. (2024), a method outputs a ranked list of findings that are more prevalent in $\mathcal{R}_A$ than in $\mathcal{R}_B$:

$$G : (\mathcal{R}_A, \mathcal{R}_B) \mapsto \mathcal{C}_{A>B} = (c_1, c_2, \ldots, c_n). \tag{1}$$

The task is hard for two reasons. The output space is open-ended, spanning pathologies ("moderate right pleural effusion"), devices and their positions ("nasogastric tube in the bronchus"), and acquisition properties ("single-view radiograph"). And cohorts contain hundreds to thousands of studies, more than any MLLM can read at once, so a method must reason about cohorts it can only sample.

3.2 RADDIFFBENCH

Stage 1: radiologists select difference concepts. We give GPT-4o (OpenAI, 2023) real MIMIC-CXR reports (Johnson et al., 2019) as context, so that its proposals use realistic clinical language, and ask it to brainstorm clinically meaningful differences between cohorts (e.g., "moderate right pleural effusion" vs. "no pleural effusion"). After a six-level deduplication (GPT-4o clustering, BM25 retrieval, and lemmatization), three radiologists rate the remaining 150 concepts on a 1–3 clinical-utility scale. We keep the 57 concepts rated 3, which the radiologists grade as easy (23), medium (21), or hard (13) based on their expertise (all pairs in Appendix B).

Stage 2: reports assign studies to cohorts. As no reliable open-vocabulary classifier exists for radiographs, we label each MIMIC-CXR study through its report, where findings are far easier to recognize than in the image: BM25 (Robertson & Zaragoza, 2009) retrieves candidate reports, and GPT-4o-mini verifies each against the concept, quoting the supporting evidence from the report (prompts in Appendix B). A radiologist audit found this labeling over 95% precise, and cohorts contain 307 images on average (Table 1). Cohorts may also differ in correlated findings (e.g., effusion

with atelectasis), so the ground truth is the *intended* difference rather than the only one, and accuracy is a conservative measure of discovery quality.

3.3 Evaluation

An LLM judge grades every prediction against the ground truth. Following Dunlap et al. (2024), GPT-4.1-nano, shown the descriptions of both cohorts, grades each c_i as a full, partial (broader or narrower), or no match, $m(c_i) \in \{1, 0.5, 0\}$, and we report

$$\mathrm{Acc@}k = \frac{1}{|\mathcal{D}|} \sum_{(\mathcal{R}_A, \mathcal{R}_B) \in \mathcal{D}} \max_{i \le k} m(c_i), \qquad k \in \{1, 5, N\}, \tag{2}$$

over the 57 pairs $\mathcal{D}$, where N is the full list (at most 100 hypotheses). LLM judges agree well with humans on such comparisons (Dunlap et al., 2024; Dubois et al., 2023); to confirm this in radiology, a radiologist independently graded 51 random predictions (17 per judge score), and their grades correlate with the judge's at Spearman $\rho = 0.785$.

4 Method

★ **Key idea.** Treat discovery as search over natural-language hypotheses: **propose** candidates from a few studies with rich radiological reasoning, **verify** each one on every study with cheap embeddings, and **refine** by feeding the verified hypotheses and zoomed-in evidence back to the proposer.

Following VisDiff (Dunlap et al., 2024), RadDiff separates generation from verification: generating hypotheses requires expensive multimodal reasoning, whereas verifying them requires only embeddings and therefore scales to entire cohorts. We first define verification (Section 4.1), then generation (Section 4.2), then the loop that connects them (Section 4.3); Algorithm 1 summarizes the procedure.

4.1 Verifying Hypotheses on the Full Cohorts

The ranker scores a hypothesis by how well it separates the two cohorts. Let $v(x, c) = \cos\big(f_{\text{img}}(x), f_{\text{txt}}(c)\big)$ be the alignment between image x and description c under a vision–language embedding model. We score c by the area under the ROC curve (AUROC) of this alignment between the cohorts:

$$s(c) \;=\; \mathrm{AUROC}\big(\{v(x,c)\}_{x \in \mathcal{R}_A}, \{v(x,c)\}_{x \in \mathcal{R}_B}\big) \;=\; \Pr\big[v(x_A, c) > v(x_B, c)\big], \tag{3}$$

the probability that a study drawn from $\mathcal{R}_A$ matches c better than one drawn from $\mathcal{R}_B$; $s(c) = 0.5$ means no difference. The AUROC is invariant to monotone rescaling of v, so scores are comparable across hypotheses, and each hypothesis costs a single text embedding regardless of cohort size. ❶ **Knowledge injection:** general-domain CLIP cannot verify findings such as "blunting of the right costophrenic angle," so RadDiff computes v with CheXzero (Tiu et al., 2022), a CLIP model trained on chest radiographs and their reports.

4.2 Proposing Hypotheses from Radiological Evidence

The proposer turns small samples into candidate findings, and what it reads determines what it can find. In each round, it samples m studies per cohort and asks an MLLM to list findings more frequent in $\mathcal{X}_A$ than in $\mathcal{X}_B$, phrased as "more *X* in group A" so that each can be verified independently. ❶ **Knowledge injection:** instead of general-domain captions, which rarely mention findings, RadDiff captions each study with the chest-radiograph foundation model CheXagent (Chen et al., 2024) and frames the task for a radiologist (Appendix C). ❷ **Multimodal reasoning:** because a caption omits whatever the captioner misses, the proposer also reads the images, as one 4×5 grid per cohort. Proposals from R rounds are pooled, deduplicated by the LLM, and verified with Eq. (3).

4.3 Refining Hypotheses in a Closed Loop

Feedback from the full cohorts turns single-pass guessing into search. A single pass over a few dozen studies is noisy, and the proposer never learns which of its hypotheses hold. ❸ **Iterative refinement:** after each iteration, the top-k hypotheses and their scores enter the proposer's prompt with an

Algorithm 1 RADDIFF. Badges mark where each component enters.

Require: cohorts $\mathcal{R}_A, \mathcal{R}_B$; iterations T; rounds R; sample size m; feedback size k

$\mathcal{F} \leftarrow \emptyset$ ▷ verified hypotheses fed back to the proposer
for $t = 1, \ldots, T$ **do**
  $\mathcal{C} \leftarrow \emptyset$
  for $r = 1, \ldots, R$ **do**
    $\mathcal{X}_A, \mathcal{X}_B \leftarrow$ sample m studies from $\mathcal{R}_A$ and from $\mathcal{R}_B$
    $\mathcal{T} \leftarrow \textsc{Caption}(\mathcal{X}_A \cup \mathcal{X}_B)$ ▷ ❶ knowledge injection: CheXagent
    if $t = 1$ **then**
      $\mathcal{V} \leftarrow \{\textsc{Grid}(\mathcal{X}_A), \textsc{Grid}(\mathcal{X}_B)\}$ ▷ ❷ multimodal reasoning
    else
      $\mathcal{B} \leftarrow \textsc{Localize}(\mathcal{F}, \mathcal{X}_A, \mathcal{X}_B)$ ▷ ❹ visual search: one box per hypothesis
      $\mathcal{V} \leftarrow \{\textsc{Grid}(\textsc{Crop}(\mathcal{X}_A \cup \mathcal{X}_B, b)) : b \in \mathcal{B}\}$
    end if
    $\mathcal{C} \leftarrow \mathcal{C} \cup \textsc{Propose}(\mathcal{T}, \mathcal{V}, \mathcal{F})$ ▷ MLLM proposer
  end for
  $\mathcal{C} \leftarrow \textsc{Dedup}(\mathcal{C})$; compute $s(c)$ for all $c \in \mathcal{C}$ with Eq. (3) ▷ ❶ knowledge injection: CheXzero
  $\mathcal{F} \leftarrow$ the k highest-scoring hypotheses in $\mathcal{C}$, with their scores ▷ ❸ iterative refinement
end for
return $\mathcal{C}$ sorted by $s(c)$

instruction to refine them, much as a radiologist revises a differential diagnosis after reviewing more cases. In effect, the proposer searches the hypothesis space with $s(c)$ as its objective. ❹ **Visual search:** small, local findings, such as a catheter tip or one blunted costophrenic angle, are lost when 20 radiographs share one grid. Inspired by guided visual search in MLLMs (Wu & Xie, 2024), from the second iteration on the proposer first returns a normalized box $(x_1, y_1, x_2, y_2) \in [0, 1]^4$ for each top-k hypothesis. Because frontal radiographs share a canonical anatomical layout, we crop the same region from every sampled study at full resolution and recompose the crops into zoomed grids, which replace the whole-image grids. Like a radiologist returning to a suspicious region, the proposer thus re-examines the studies it has already read, now at the scale of the finding. After T iterations, RADDIFF returns the hypotheses ranked by $s(c)$.

5 EXPERIMENTS

Setup. RADDIFF uses CheXagent-8b (Chen et al., 2024) as the captioner, GPT-4.1-nano as the proposer, and CheXzero (Tiu et al., 2022) as the ranker, with $m = 20$ studies per cohort, $R = 3$ rounds of up to 10 hypotheses, $T = 2$ iterations, $k = 5$ feedback hypotheses, and a single seed (prompts in Appendix C). We start from VisDiff (Dunlap et al., 2024), with its general-domain captioner and CLIP ranker, and add RADDIFF's components one at a time. As an oracle reference, the proposer reads the ground-truth reports instead of captions; since RADDIFFBENCH cohorts were defined from these reports, this setting leaks labels and only gauges the value of better captions.

5.1 MAIN RESULTS

RADDIFF turns a failing baseline into a usable discovery tool. VisDiff reaches 1.8% Acc@1 and 28.9% Acc@N (Table 2): general-domain captions rarely mention radiological findings, and general-domain CLIP cannot verify them. RADDIFF reaches 47.4% Acc@1 (+45.6 points) and 90.4% Acc@N, improving at every difficulty level (60.9%, 47.6%, and 23.1% Acc@1 on easy, medium, and hard pairs). The gain far exceeds the uncertainty of a 57-pair benchmark: approximate 95% Wilson intervals for Acc@1 are 35.0–60.1% for RADDIFF and 0.3–9.3% for VisDiff. Since one pair corresponds to 1.75 points, however, we focus on trends that are consistent across metrics when comparing components.

Ranking, not generation, is now the bottleneck. Of the 52.6 points of Acc@1 that RADDIFF misses, 43.0 come from intended differences that are proposed but ranked too low (Acc@N is 90.4%), and only 9.6 from differences that are never proposed. Stronger verification, e.g., with better medical embedding models, is thus a direct path to further gains.

Table 2: Results on RADDIFFBENCH. Acc@k (%; Eq. (2)) overall and by difficulty (number of pairs in parentheses); bars show overall Acc@1. Starting from VisDiff (Dunlap et al., 2024), components ❶–❹ are added cumulatively; the two refinement rows are alternative feedback sizes, and the full model adds visual search to top-5 feedback. **Bold**: best; underlined: second best, excluding the gray oracle rows, which read the ground-truth reports.

Component	Variant	**Overall** (57) @1	@5	@N	**Easy** (23) @1	@5	**Medium** (21) @1	@5	**Hard** (13) @1	@5
Baseline	VisDiff	1.8	3.5	28.9	4.3	4.3	0.0	2.4	0.0	3.8
❶ Knowledge injection	+ CheXagent captioner	9.6	30.7	78.1	17.4	50.0	2.4	21.4	7.7	11.5
	+ CheXzero ranker	28.9	57.9	78.1	41.3	65.2	28.6	73.8	7.7	19.2
	+ radiology prompt	29.8	62.3	84.2	32.6	65.2	40.5	**78.6**	7.7	30.8
❷ Multimodal reasoning	+ image grids	33.3	63.2	86.8	45.7	71.7	40.5	**78.6**	0.0	23.1
❸ Iterative refinement	+ top-5 feedback	43.9	**69.3**	89.5	58.7	73.9	45.2	73.8	15.4	**53.8**
	+ top-10 feedback	45.6	65.8	86.0	50.0	76.1	**57.1**	76.2	19.2	30.8
❹ Visual search	RADDIFF (full)	**47.4**	67.5	**90.4**	**60.9**	**78.3**	47.6	**78.6**	**23.1**	30.8
Oracle	*GT reports as input*	*37.7*	*78.1*	*97.4*	*39.1*	*69.6*	*47.6*	*95.2*	*19.2*	*65.4*
	+ refinement	*50.9*	*78.9*	*91.2*	*52.2*	*76.1*	*54.8*	*90.5*	*42.3*	*65.4*

Table 3: Iterative refinement in action on the hard pair $\mathcal{R}_A$: *normal heart size* vs. $\mathcal{R}_B$: *cardiac silhouette remains enlarged*. First-round distractors outrank the ground truth (highlighted); conditioned on them and their scores $s(c)$, refinement converges on it.

	Hypothesis (more frequent in $\mathcal{R}_A$)	$s(c)$
Iteration 1 (top 3)	More cases with normal pulmonary vasculature	0.846
	More cases showing normal osseous structures	0.844
	More instances of **normal heart size** and mediastinal/hilar contours	0.830
Refined	**Normal heart size** and mediastinal/hilar contours in group A	
	More consistent normal findings across group A	

Better reasoning matters more than better captions. A single pass over the oracle reports reaches 37.7% Acc@1, below RADDIFF with generated captions; adding refinement to the oracle yields 50.9%, only 3.5 points above RADDIFF. Once hypotheses are verified and refined, caption quality is no longer the main bottleneck.

Hard pairs remain open. They involve subtle findings (a nodule, a displaced rib fracture) or changes relative to a prior study (worsening consolidation) that one radiograph cannot fully show. Most failures are plausible but misdirected rather than hallucinated: salient confounders such as chest tubes draw attention away from subtle pathology (40% of hard-pair errors), focal lesions lose to diffuse patterns (30%), or a sign such as an effusion is found without its cause (20%; Appendix D.1).

5.2 COMPONENT ANALYSIS

❶ **Knowledge injection gives the largest gain.** CheXagent captions raise Acc@N from 28.9% to 78.1%, but general-domain CLIP ranks the right hypotheses too low (9.6% Acc@1); CheXzero raises Acc@1 to 28.9% at unchanged Acc@N, so verification needs domain knowledge as much as generation does. Radiology-specific instructions add 6.1 points of Acc@N.

❷ **Multimodal reasoning fills gaps in the captions.** Reading images alongside captions raises Acc@1 from 29.8% to 33.3%, mainly on easy pairs (32.6% to 45.7%), whose devices and tubes are visible but inconsistently captioned. On hard pairs, whole images do not help (7.7% to 0.0% Acc@1), likely because subtle findings are hard to see in a 20-image grid, motivating visual search.

❸ **Iterative refinement gives the second-largest gain.** Top-5 feedback raises Acc@1 to 43.9% and more than doubles hard-pair Acc@5, from 23.1% to 53.8%, by promoting the ground truth over first-round distractors (Table 3); gains plateau after two to three iterations (Figure 4). The feedback size matters less (top-10: 45.6% Acc@1; top-5: 69.3% Acc@5).

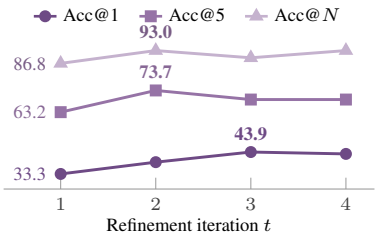


Figure 4: Accuracy (%) across refinement iterations (top-5 feedback, no visual search). Acc@5 and Acc@N peak after the second iteration and Acc@1 after the third (bold); later iterations mostly recycle earlier hypotheses.

4 **Visual search localizes the evidence.** Added to top-5 refinement, it raises Acc@1 from 43.9% to 47.4% and hard-pair Acc@1 from 15.4% to 23.1%, and yields the best overall Acc@N, with Acc@5 within about one pair. Its main benefit is interpretability: the zoomed regions converge on the relevant anatomy, such as the right lower lung for a right pleural effusion (Appendix D.2); the boxes in Figures 1 and 5–7 show the regions selected for each top hypothesis.

✓ **Takeaways.** 1 Domain knowledge makes findings both proposable and verifiable (Acc@1 from 1.8% to 29.8%); 2 images fill the gaps in captions (33.3%); 3 feedback from the full cohorts turns proposals into search (43.9%); and 4 visual search grounds each hypothesis in the region it describes (47.4%).

6 Automated Discovery in Practice

We now use RadDiff to answer Q1–Q3, which illustrate three uses: generating prognostic hypotheses, characterizing disease phenotypes, and auditing a medical AI model. For each question, we build cohorts from MIMIC-CXR (Johnson et al., 2019) linked to the clinical records of MIMIC-IV (Johnson et al., 2023) and run RadDiff in both directions. Two radiologists reviewed the top-ranked hypotheses and confirmed that they are meaningful and clinically consistent. We treat these findings as data-driven hypotheses, not causal or clinically validated conclusions.

6.1 Q1 What Distinguishes Pneumonia Non-Survivors?

Question. Hospitalized pneumonia patients vary widely in severity, and identifying high-risk patients early enables timely intervention. Deep networks predict 30-day mortality from chest radiographs (Kim et al., 2023), but the imaging features behind such predictions remain unclear: which radiographic patterns distinguish patients who die?

Setup. We compare pneumonia patients (ICD-9/10; Appendix E) who died in hospital ($\mathcal{R}_A$; 11,861 images) with those who survived at least one year ($\mathcal{R}_B$; 10,351 images), sampling equally within sex and age strata before deduplication. We repeat the analysis for 30- and 90-day mortality.

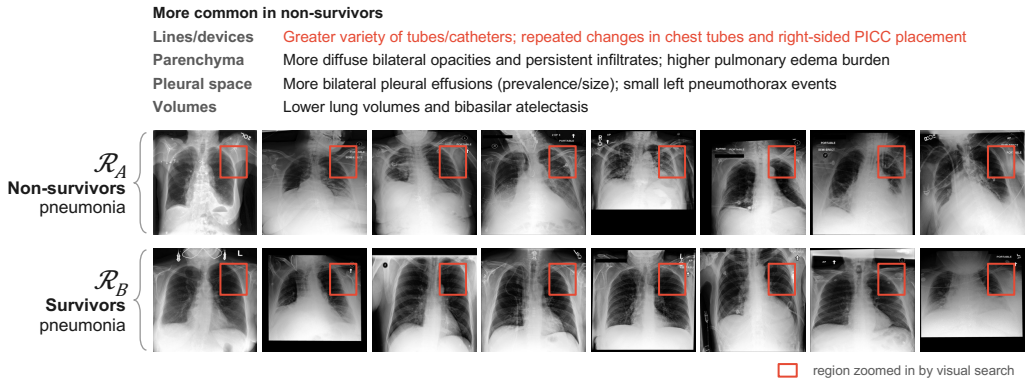


Figure 5: Q1 Pneumonia non-survivors vs. survivors. Non-survivors show more extensive disease and more interventions. Red boxes mark the regions visual search selected for the top hypothesis, "more documented changes in thoracic catheters and lines."

Findings. Non-survivors are distinguished first by the intensity of care: a greater number and variety of thoracic devices, such as peripherally inserted central catheters (PICCs) and "central venous catheters and endotracheal tubes projecting above the carina" (Figure 5), reflecting more

intensive support and more severe respiratory compromise. Beyond devices, RADDIFF finds more extensive disease, such as "extensive bilateral pulmonary opacities," "pulmonary vascular congestion," "enlarged cardiomediastinal silhouette with associated pleural effusions," and "bibasilar atelectasis," in line with guidelines that list multilobar infiltrates among the criteria for severe pneumonia (Metlay et al., 2019). For 30- and 90-day mortality, device findings fade while effusions, opacities, and hyperinflation dominate (Appendix Table 9), suggesting that chronic lung disease matters more for medium-term risk.

Discovery Q1 In-hospital death is marked by device burden and extensive bilateral disease; as the horizon lengthens, the signal shifts from acute interventions toward effusions and chronic hyperinflation.

6.2 Q2 How Does COVID-19 Present with Age?

Question. Age is a major determinant of COVID-19 severity and trajectory and may alter how the infection appears on chest radiographs, yet the radiographic features that distinguish older from younger patients have not been systematically characterized.

Setup. We compare COVID-19 patients (ICD-10) older than 60 ($\mathcal{R}_A$; 901 images) with those aged 40 or younger ($\mathcal{R}_B$; 807 images), matched on sex; the top differences agree in both directions.

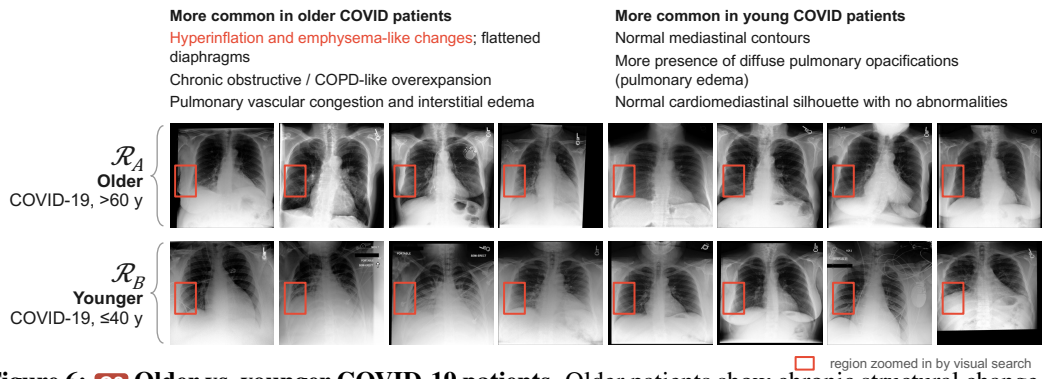


Figure 6: Q2 Older vs. younger COVID-19 patients. Older patients show chronic structural change, such as hyperinflation and emphysema-like features, whereas younger patients show more acute opacities. Red boxes mark the regions selected for "more frequent mention of hyperinflation and emphysema-like features."

Findings. Older patients show chronic structural change: "lung hyperinflation resembling chronic obstructive pulmonary disease (COPD)," "emphysema-like lucencies," flattened diaphragms, and more "vascular congestion and interstitial edema" (Figure 6). Younger patients instead show "more acute diffuse opacities." In older patients, the infection is thus superimposed on chronic cardiopulmonary disease, consistent with CT evidence that ground-glass opacities become less and effusions more prevalent with age (Okoye et al., 2022). Such age-specific phenotypes can inform age-aware triage and the evaluation of COVID-19 models across age groups.

Discovery Q2 The same infection looks different with age: older patients' radiographs are dominated by chronic COPD-like change and congestion, younger patients' by acute opacities.

6.3 Q3 What Does a Race Classifier See?

Question. Deep networks predict self-reported race from chest radiographs with high accuracy, although clinicians cannot (Gichoya et al., 2022; Lotter, 2024). Which cues do they exploit? Our goal is not to assert biological differences between groups but to expose *confounders*.

Setup. We fine-tune a DeiT vision transformer (Touvron et al., 2021) on 15,000 MIMIC-CXR radiographs (5,000 each of White, Asian, and Black patients; ∼75% validation accuracy, chance 33%) and compare the studies it assigns to White ($\mathcal{R}_A$; 662 images) or Asian patients ($\mathcal{R}_B$; 1000 images) with probability above 0.95. RADDIFF thus audits the classifier rather than the populations.

Findings. The separating cues are procedural and contextual, not anatomical. The White subset has more devices in specific positions, such as "Port-A-Cath placements terminating at the cavoatrial

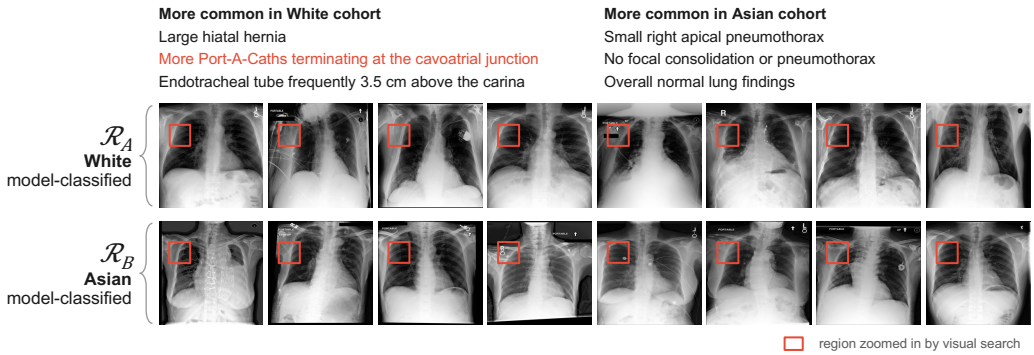


Figure 7: Q3 Radiographs a fine-tuned vision transformer classifies as White vs. Asian with high confidence. Devices, acquisition context, and apparent normality, rather than anatomy, separate the two subsets. Red boxes mark the regions selected for the top hypothesis, "more cases with the tip of vascular access devices (Port-A-Cath) terminating at the cavoatrial junction."

junction" and "endotracheal tubes about 3.5 cm above the carina" (Figure 7), and more abnormalities such as "large hiatal hernia." The Asian subset is instead characterized by normality: three of its top five and seven of its top ten hypotheses describe normal findings (e.g., "well-expanded lungs"), versus none for the White subset. The classifier's confident predictions thus entangle race with apparent disease severity and acquisition, consistent with prior work on acquisition factors (Lotter, 2024) and underdiagnosis (Seyyed-Kalantari et al., 2021); White vs. Black subsets show the same pattern (Appendix E). By surfacing such confounders in plain language, RADDIFF can inform dataset rebalancing, acquisition standardization, and fairness-aware model design.

Discovery Q3 A race classifier's most confident predictions track devices, acquisition context, and how normal a study looks rather than anatomy, entangling predicted race with apparent disease severity.

7 CONCLUSION

We introduced RADDIFF, an agent that automates radiology discovery by proposing, verifying, and refining natural-language hypotheses about how two cohorts differ, and RADDIFFBENCH, a radiologist-validated benchmark for this task. RADDIFF raises top-1 accuracy from 1.8% to 47.4% and yields clinically coherent discoveries about pneumonia mortality, age-dependent COVID-19 phenotypes, and the shortcuts of a race classifier. More broadly, pairing an expensive reasoner that reads a few samples with a cheap verifier that scores every study is an effective recipe for open-ended discovery when no model can read the whole dataset.

Limitations and outlook. RADDIFF currently handles 2D chest radiographs; RADDIFFBENCH derives its cohorts from reports, so reporting errors propagate into its labels, and its 57 pairs make differences of one or two pairs hard to resolve. RADDIFF generates hypotheses, not conclusions: its findings reflect correlations in the data, including confounders such as care intensity, and require expert review and formal statistical testing before they inform practice (Appendix A). Extending RADDIFF to 3D imaging and coupling it with hypothesis testing are natural next steps toward AI systems that help clinicians ask, and answer, population-level questions of their imaging archives.

AI USE STATEMENT

Large language models are components of the method and benchmark studied in this paper, and their roles are described in the main text: GPT-4o and GPT-4o-mini propose difference concepts and match reports during benchmark construction (Section 3); GPT-4.1-nano is RADDIFF's proposer and the evaluation judge; and CheXagent generates captions (Section 4). All benchmark concepts were validated by radiologists, and the LLM judge was validated against a radiologist (Section 3.3). Generative AI tools were also used to assist coding, and edit the manuscript for clarity, consistency, and formatting. The authors reviewed and revised all AI-assisted content and take full responsibility for this paper, including its text, claims, and artifacts.

Ethics Statement

This work uses MIMIC-CXR and MIMIC-IV, de-identified datasets that we accessed under the PhysioNet credentialed data use agreement after completing the required human-subjects training; no new patient data were collected. The radiologists who validated the benchmark and reviewed the findings are co-authors of the paper. Our race analysis uses self-reported race, a social rather than biological category; we study it to audit what a classifier relies on, and we make no claims about biological differences between groups. RadDiff generates hypotheses for expert review. It is not a diagnostic device, and its outputs, which can reflect spurious correlations in the data, should not inform individual clinical decisions without clinical validation.

Reproducibility Statement

Section 4 and Algorithm 1 describe RadDiff, Section 5 lists the models and hyperparameters, and Appendices B and C contain every prompt used for benchmark construction, hypothesis proposal, iterative refinement, visual search, and evaluation. The supplementary material contains anonymized code, configuration files. We will release the full code and cached model outputs that reproduce Table 2. RadDiffBench is distributed as MIMIC-CXR image identifiers with cohort assignments and ground-truth descriptions, so that anyone with PhysioNet credentials can rebuild it, and the application cohorts are defined by the ICD codes in Appendix E.

References

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

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision (ECCV)*, pp. 1–21, 2022.

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

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24199–24208, 2024.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4 (6):e406–e414, 2022.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3942–3951, 2021.

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

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Changi Kim, Eui Jin Hwang, Ye Ra Choi, Hyewon Choi, Jin Mo Goo, Yisak Kim, Jinwook Choi, and Chang Min Park. A deep learning model using chest radiographs for prediction of 30-day mortality in patients with community-acquired pneumonia: development and external validation. *American Journal of Roentgenology*, 221(5):586–598, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

William Lotter. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15(1):7465, 2024.

Joshua P. Metlay, Grant W. Waterer, Ann C. Long, Antonio Anzueto, Jan Brozek, Kristina Crothers, Laura A. Cooley, Nathan C. Dean, Michael J. Fine, Scott A. Flanders, Marie R. Griffin, Mark L. Metersky, Daniel M. Musher, Marcos I. Restrepo, and Cynthia G. Whitney. Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. *American Journal of Respiratory and Critical Care Medicine*, 200(7):e45–e67, 2019.

Chukwuma Okoye, Panaiotis Finamore, Giuseppe Bellelli, Alessandra Coin, Susanna Del Signore, Stefano Fumagalli, Pietro Gareri, Alba Malara, Enrico Mossello, Caterina Trevisan, et al. Computed tomography findings and prognosis in older COVID-19 patients. *BMC Geriatrics*, 22(1):166, 2022.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 38154–38180, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pp. 10347–10357, 2021.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, pp. 58–76, 2024.

Penghao Wu and Saining Xie. V^*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference (MLHC)*, pp. 2–25, 2022.

Appendix overview. Appendix A discusses limitations. Appendix B details RADDIFFBENCH: all 57 cohort pairs, the prompts used to construct and evaluate it, and examples. Appendix C lists the prompts used by RADDIFF. Appendix D reports additional results: a failure analysis, qualitative examples of iterative refinement and visual search, an end-to-end case study, and further ablations. Appendix E gives details and extended results for the case studies of Section 6.

A LIMITATIONS

Scope. RADDIFF currently handles 2D chest radiographs. The loop itself is modality-agnostic, but 3D modalities such as CT and MRI will require volumetric captioners, rankers, and visual search.

Benchmark. Cohorts in RADDIFFBENCH are defined from radiology reports, so reporting errors and omissions propagate into the cohorts. Each pair is defined by a single intended difference, although cohorts may differ in correlated findings, so Acc@k is a conservative measure of discovery quality. Several hard concepts are defined relative to a prior study (e.g., "worsening consolidation"), which a single radiograph cannot fully reveal. With 57 pairs, one pair corresponds to 1.75 points, so small differences between variants should be interpreted with care; all results use a single seed.

Evaluation. Accuracy is graded by an LLM judge. The judge agrees well with a radiologist (Spearman $\rho = 0.785$), but it can still misgrade borderline cases, especially partial matches.

Robustness of the loop. On hard pairs, errors in localization or ranking can propagate through iterative refinement, and later iterations can recycle earlier hypotheses rather than explore new ones (Appendix D.4).

Use of discoveries. RADDIFF generates hypotheses, not conclusions. Its findings reflect correlations in the data, including confounders such as care intensity, and they require expert review and formal statistical testing before informing research conclusions or clinical practice. It is designed to support, not to replace, radiologists.

B RADDIFFBENCH DETAILS

B.1 ALL COHORT PAIRS

Table 4 lists the 57 radiologist-validated cohort pairs of RADDIFFBENCH, grouped by difficulty.

B.2 CONSTRUCTION PROMPTS

Figures 8, 9, and 10 show the prompts used to propose difference concepts from sampled reports, to deduplicate them, and to assign reports to cohorts.

Stage 1 • Difference concept proposal

```
List all hypothetical potential differences between sets of chest x-ray radiology scans.
These could include but not limited to variations in tissue density, presence of
abnormalities such as tumors,
lesions, or fractures, and any noticeable changes in anatomical structures.
Give me exactly {num_differences} differences in the format of A vs B in a JSON file.
Store condition A and B in seperate fields in the JSON. The JSON format should be of the
following:
[
{ "condition_A":  "*insert condition A*", "condition_B":  "*insert condition B*" },
...
]
Ensure these distinctions reflect the detailed nuances characteristic of radiology
reports.
They should not be broad classification differences but rather subtle, intricate
variations.
Here are sample radiology reports to help you:
{sample_reports}
```

Figure 8: Prompt used to propose candidate difference concepts from sampled MIMIC-CXR reports (GPT-4o).

Table 4: All 57 cohort pairs of RADDIFFBENCH, grouped by radiologist-assigned difficulty. Concept strings are lightly copy-edited for readability (e.g., undoing lemmatization introduced during deduplication); the released benchmark files contain the original strings.

#	Cohort A ($\mathcal{R}_A$)	Cohort B ($\mathcal{R}_B$)
Easy	23 pairs	
1	right subclavian central venous catheter present	no subclavian central venous catheter observed
2	previously noted pulmonary edema resolved	moderate pulmonary edema with slightly improved aeration
3	mild to moderate cardiomegaly	moderate cardiomegaly, mildly stable
4	NG tube in bronchus	NG tube in stomach
5	unstable cardiomegaly with pulmonary edema	stable cardiomegaly, no edema
6	abnormal chest radiograph	normal chest radiograph
7	left apical pleural tube in place	no pleural tube
8	enteric tube terminating below the diaphragm	no enteric tube
9	poc catheter tip in the low SVC	poc catheter not visualized
10	enlarged heart size	normal heart size
11	single-view chest radiograph	PA and lateral chest radiograph
12	enlarged cardiac silhouette	normal cardiac silhouette
13	patchy middle lobe opacity	clear middle lobe
14	nasogastric tube removed	nasogastric tube present
15	hyperinflated lungs with flattened diaphragms	clear lungs
16	PICC terminating in the low SVC	PICC absent or not in the low SVC
17	mild cardiomegaly	significant cardiomegaly
18	opacity concerning for pneumonia	no evidence of pneumonia
19	clear lungs without consolidation, effusion, or pneumothorax	right lower lobe pneumonia
20	interstitial abnormality with vascular congestion	no interstitial abnormality or congestion
21	residual patchy opacity in the right lung	clear right lung
22	hyperinflated lungs	normal lung inflation
23	moderate to large right pneumothorax	no pneumothorax
Medium	21 pairs	
24	moderate right pleural effusion	no pleural effusion
25	hypoinflated lungs with perihilar opacity	well-inflated, clear lungs
26	small bilateral pleural effusions	no pleural effusion or pneumothorax
27	small to moderate left pneumothorax	no pneumothorax
28	bilateral pneumothorax	no pneumothorax
29	clear lungs	diffuse interstitial opacity
30	right middle lobe pneumonia	no pneumonia
31	subtle opacity at the left lung base	clear lung base
32	moderate to severe cardiomegaly	normal heart size
33	well-inflated, clear lungs	bibasal interstitial opacity
34	normal heart size	mildly to moderately enlarged heart
35	new opacity in the left mid/lower lung	no new opacity
36	dense right upper lobe consolidation	no consolidation
37	high sensitivity for pneumothorax	low sensitivity for pneumothorax (supine)
38	lungs mostly clear	bibasilar opacity and lung mass
39	rightward mediastinal shift	no mediastinal shift
40	right apical opacity	no apical opacity
41	small bilateral pleural effusions	no effusion
42	signs of tuberculosis infection	no evidence of tuberculosis
43	moderate edema and effusion	minimal edema, no effusion
44	low lung volumes and bibasilar opacity	normal lung volumes, clear lungs
Hard	13 pairs	
45	displaced rib fracture	no displaced rib fracture
46	stable airspace consolidation	worsening airspace consolidation
47	confluent left perihilar opacity	clear perihilar region
48	elevated pulmonary venous pressure	normal pulmonary venous pressure
49	low lung volumes	normal lung volumes
50	esophageal perforation	no perforation
51	normal heart size	cardiac silhouette remains enlarged
52	pulmonary nodule	no pulmonary nodule
53	clear basal parenchyma	basal atelectasis
54	worsening retrocardiac opacification	no significant change
55	multilevel degenerative changes of the spine	normal spinal structure
56	mildly enlarged, unchanged heart	heart not enlarged
57	moderate cardiomegaly	normal cardiac silhouette

Stage 1 • Concept de-duplication

```
Below are hypothetical differences between chest X ray.  For the below set of differences,
remove any differences that are semantically and medically similar to each other.
Please be sure to tell me which differences were removed and explain your reasoning.
{differences}
Return the final differences, with duplicates removed, as a JSON in the following format:
{
differences:  [
{
"condition_A":  "",
"condition_B":  "",
},
...
]
}
```

Figure 9: Prompt used to remove semantically duplicated difference concepts.

Stage 2 • Report-based cohort assignment

```
We have the following condition of the format A vs B respectively:  {difference}.
Given the following {len(reports)} radiology reports, group each report into either having
condition A or B or neither.
Classify each report into only one group exactly.  Do not place a report in multiple
groups.
Provide reasoning and direct evidence in quotes from the report to justify each grouping.
Put the final output in a JSON with the following format:
{
"group A":  [
{
"report_index":  "",
"reasoning":  "",
"direct_evidence":  "",
},
...
],
"group B":  [
{
"report_index":  "",
"reasoning":  "",
"direct_evidence":  "",
},
...
],
"neither":  [
{
"report_index":  "",
"reasoning":  "",
"direct_evidence":  "",
},
...
]
}
Please make sure to classify ALL the reports shown below:
{reports}
```

Figure 10: Prompt used to assign each retrieved report to cohort A, cohort B, or neither (GPT-4o-mini).

B.3 Evaluation Prompt

Figure 11 shows the prompt of the LLM judge (GPT-4.1-nano). Scores of 2, 1, and 0 correspond to full, partial, and no match, i.e., $m(c) \in \{1, 0.5, 0\}$ in Eq. (2).

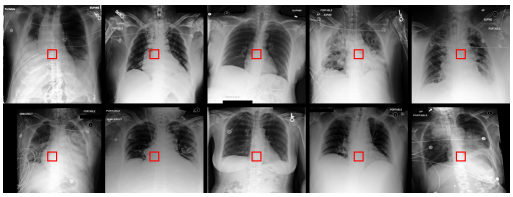
Evaluation • LLM judge

```
I am a machine learning researcher working on identifying differences between two groups
of radiology images.  My goal is to determine how well a given prediction corresponds to
the findings or conditions that are more commonly present in Group A compared to Group B.
You will receive descriptions of Group A and Group B, along with a prediction.
Your task:
Evaluate whether the prediction is more aligned with Group A or Group B, using the
following scoring system:
2:  Fully aligned with Group A.
1:  Partially aligned with Group A (i.e., the prediction is closer to Group A than Group B
but represents a broader or narrower concept).
0:  Not aligned with Group A (i.e., more aligned with Group B or represents a completely
different concept).
Reference Example 1:
Group A: "Left-sided opacity" and Group B: "Right-sided opacity"
Prediction:  "Left-sided opacity" → Score:  2 (fully aligned with Group A)
Prediction:  "Left lung consolidation" → Score:  2 (fully aligned with Group A)
Prediction:  "Unilateral lung opacity" → Score:  1 (broader but closer to Group A)
Prediction:  "Right-sided opacity" → Score:  0 (aligned with Group B)
Reference Example 2:
Group A: "Pleural effusion" and Group B: "No pleural effusion"
Prediction:  "Pleural effusion" → Score:  2 (fully aligned with Group A)
Prediction:  "Fluid in the pleural space" → Score:  2 (fully aligned with Group A)
Prediction:  "Increased fluid in the chest cavity" → Score:  1 (broader but closer to
Group A)
Prediction:  "Normal lungs" → Score:  0 (aligned with Group B)
Now, analyze the following using similar reasoning from the above examples as a guide.
Group A: {gt_a}
Group B: {gt_b}
Prediction:  {hypothesis}
Please respond with 2, 1, or 0, based on the alignment of the prediction with Group A.
```

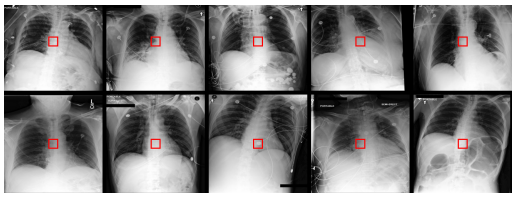
Figure 11: Prompt used by the LLM judge to score a predicted difference against the ground-truth descriptions of $\mathcal{R}_A$ and $\mathcal{R}_B$.

B.4 EXAMPLES

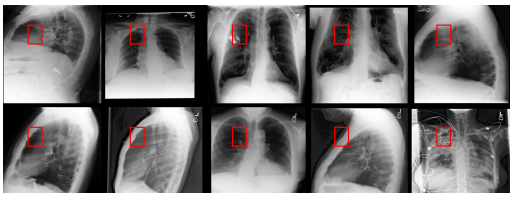
Figures 12, 13, and 14 show three pairs from each difficulty level, together with the regions selected by visual search and the top differences proposed by RADDIFF.



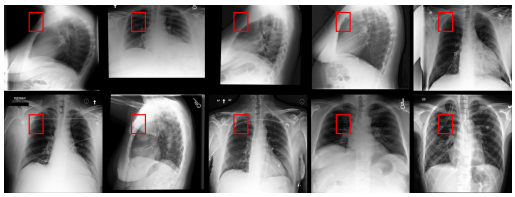
$\mathcal{R}_A$: NG tube in bronchus.



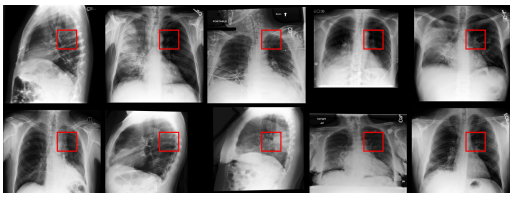
$\mathcal{R}_B$: NG tube in stomach.



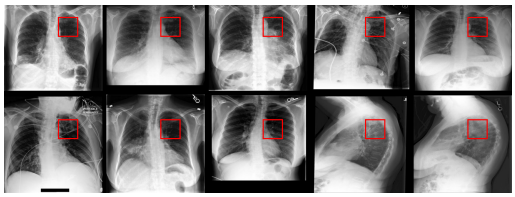
$\mathcal{R}_A$: Hyperinflated lung with flattening of the hemidiaphragm.



$\mathcal{R}_B$: Clear lung.

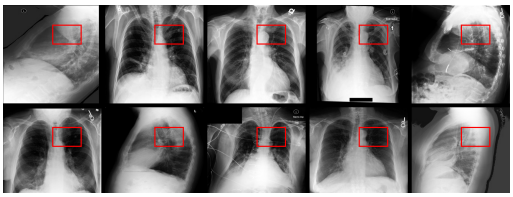


$\mathcal{R}_A$: Patchy middle lobe opacity.

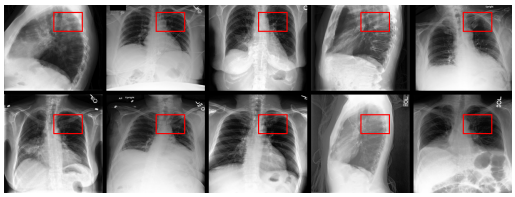


$\mathcal{R}_B$: Clear middle lobe.

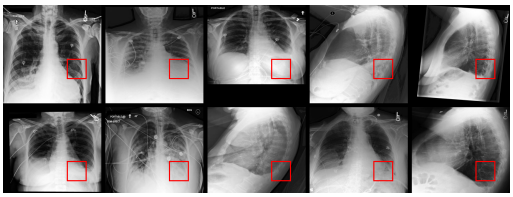
Figure 12: Easy examples. RADDIFF localizes salient regions and surfaces clinically meaningful cohort-level differences. Top row: visual search focuses on the carina region, producing predictions such as "Higher frequency of endotracheal tubes located above the carina in Group A." Middle row: lung hyperinflation, producing "More instances of hyperinflated lungs without focal consolidation or effusion in Group A." Bottom row: RADDIFF proposing differences such as "More consolidation and opacity patterns suggestive of pneumonia in Group A."



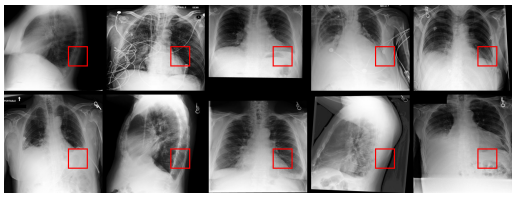
$\mathcal{R}_A$: Pneumonia in right middle lobe.



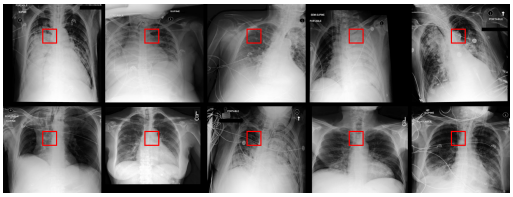
$\mathcal{R}_B$: No pneumonia.



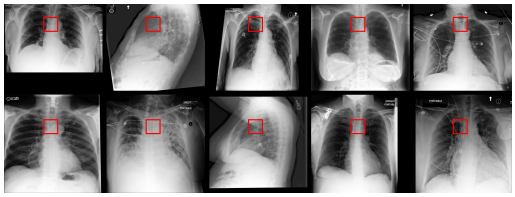
$\mathcal{R}_A$: Bilateral small pleural effusions.



$\mathcal{R}_B$: No pleural effusion or pneumothorax.

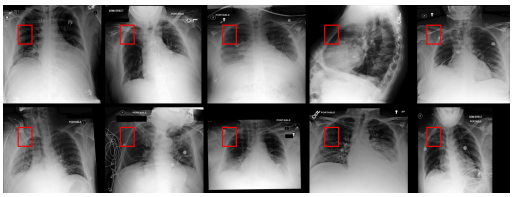


$\mathcal{R}_A$: Dense right upper lobe airspace consolidation.

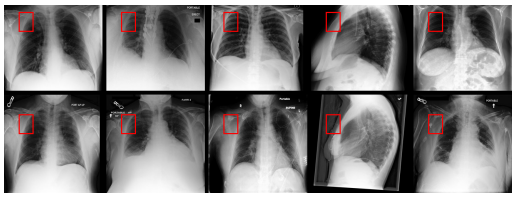


$\mathcal{R}_B$: No airspace consolidation.

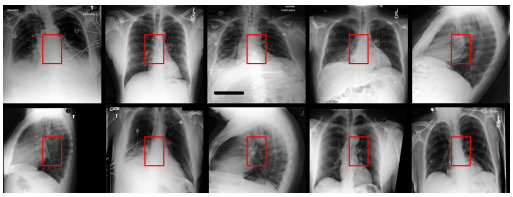
Figure 13: Medium examples. We show the set names and the top two differences generated by RADDIFF. Top row: "More frequent right lower lobe consolidations suggestive of pneumonia in Group B" and "Multifocal pneumonia with opacities in multiple lobes in Group A." Middle row: "Presence of pleural effusions with atelectasis and consolidations in Group A" and "Bilateral pleural effusions obscuring hemidiaphragms in Group A." Bottom row: "More extensive bilateral parenchymal opacities in Group A" and "Presence of large right pleural effusion with adjacent atelectasis/consolidation in Group A."



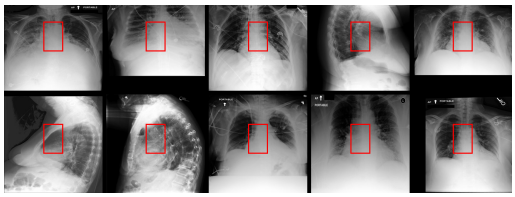
$\mathcal{R}_A$: Elevated pulmonary venous pressure.



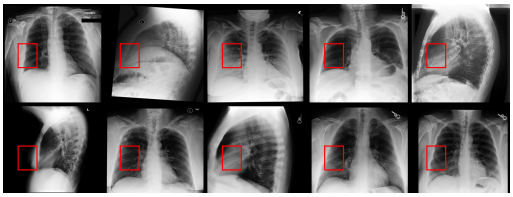
$\mathcal{R}_B$: Normal pulmonary venous pressure.



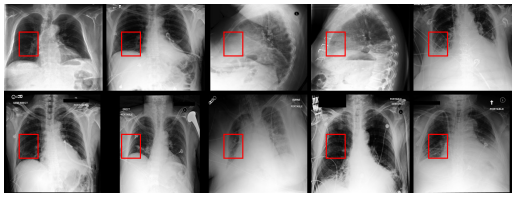
$\mathcal{R}_A$: Normal heart size.



$\mathcal{R}_B$: Enlarged cardiac silhouette.



$\mathcal{R}_A$: Clear basal parenchyma.



$\mathcal{R}_B$: Basal atelectasis.

Figure 14: Hard examples. We show the set names and the top two differences generated by RADDIFF. Top row: "More extensive pleural effusions with underlying consolidation in Group A" and "Pulmonary edema and fluid overload." Middle row: "More cases with normal heart size and mediastinal/hilar contours in Group A" and "Absence of hyperinflation and diaphragm flattening in Group A." Bottom row: "Normal cardiomediastinal silhouette and well-expanded lungs in Group A" and "Basal atelectasis with associated abnormalities in Group B."

C RADDIFF PROMPTS

We list the prompts used for multimodal proposal (Figure 15), iterative refinement (Figure 16), and visual search (Figures 17 and 18). Captions are generated by CheXagent with the prompt "Describe this image in detail."

Proposer • Multimodal proposal (iteration 1)

```
The following are the results of captioning two groups of chest X-ray images used for a
detailed medical analysis:
{text}
We also have the two groups of medical chest X-ray images shown below as well.  Group A
chest X-rays are shown in the first image, while Group B Chest X-rays are part of the
second image.
Your task:
You are the best radiologist in the world.  Can you identify the most salient differences
between these two groups of chest X-rays, using the above captions and attached images.
Provide the differences in a clear way (i.e "A has more xxx", but only return "xxx")
Make sure to analyze the captions and images carefully and extract 5-10 salient
differences that are more frequently observed in Group A compared to Group B.
Make sure to only provide information of what group A has more of.
Don't mention anything about group B in your set of differences.
Answer with a list of the most distinct salient differences:
```

Figure 15: Prompt used by the proposer in the first iteration, with CheXagent captions and one image grid per cohort.

Proposer • Iterative refinement

```
The following are the results of captioning two groups of chest X-ray images used for a
detailed medical analysis:
{text}
We also have the two groups of medical chest X-ray images shown below as well.  Group A
chest X-rays are shown in the first image, while Group B Chest X-rays are part of the
second image.
Your task:
You are the best radiologist in the world.  Can you identify the most salient differences
between these two groups of chest X-rays, using the above captions and attached images.
Provide the differences in a clear way (i.e "A has more xxx", but only return "xxx")
Make sure to analyze the captions and images carefully and extract 5-10 salient
differences that are more frequently observed in Group A compared to Group B.
Make sure to only provide information of what group A has more of.
Don't mention anything about group B in your set of differences.
Here are the top {top} differences and scores from the previous round:
{prev_results}
Refine and improve upon these results.
Answer with a list of the most distinct salient differences:
```

Figure 16: Prompt used for iterative refinement: the top-k verified hypotheses and their scores are fed back to the proposer.

Visual search • Region localization

```
The following are the results of captioning two groups of chest X-ray images used for a
detailed medical analysis:
{text}
We also have the two groups of medical chest X-ray images shown below as well.  Group A
chest X-rays are shown in the upper half of the image, while Group B Chest X-rays are part
of the lower half of the image.
Here are the top {top} differences and scores from the previous round:
{prev_results}
For each of the top {top} findings listed below, we'd like you to pick one area on a chest
X-ray image that best shows the difference.
```

```
Please give us a set of four numbers - x1, y1, x2, y2 - that describe a rectangle covering
that area.  Each number should be between 0 and 1, and they should be based on the size
of the image (for example, 0 means the far left or top of the image, and 1 means the far
right or bottom).  We'll use these rectangles to crop the images and take a closer look at
the areas where the differences are most visible and clinically important.
```

Figure 17: Prompt used by visual search to localize, for each top-k hypothesis, the region that best shows the difference.

Visual search • Proposal from zoomed regions

```
MEDICAL CONTEXT: You are analyzing two distinct cohorts of chest X-ray images for
differential diagnostic patterns.
CAPTION ANALYSIS DATA:
{text}
VISUAL DATA: The attached images show 5 cropped regions highlighting previously identified
differences.  Each image has:
- UPPER SECTION (Group A): Separated by a visual gap from Group B
- LOWER SECTION (Group B): Below the visual gap
CLINICAL TASK:
As a board-certified radiologist, perform comparative analysis to identify radiological
findings that are statistically more prevalent in Group A.
ANALYSIS REQUIREMENTS:
1.  Focus on specific anatomical structures and pathological findings
2.  Use precise medical terminology (e.g., "consolidation," "pleural effusion,"
"cardiomegaly")
3.  Consider both caption data and visual evidence
4.  Prioritize clinically significant differences
PREVIOUS ITERATION RESULTS:
{prev_results}
REFINEMENT INSTRUCTIONS:
- Enhance specificity of previous findings
- Eliminate false positives or artifacts
- Focus on reproducible patterns across multiple images
- Prioritize diagnostically relevant features
OUTPUT FORMAT:
Provide exactly 5-10 refined findings as single-phrase medical terms (e.g., "bilateral
lower lobe consolidation", "enlarged cardiac silhouette", "pleural thickening"):
```

Figure 18: Prompt used to propose refined differences from the zoomed-in grids produced by visual search.

D Additional Results

D.1 Failure Analysis

To understand where RadDiff fails, we analyze its errors on the hard subset of RadDiffBench (Table 5). Most failures are clinically plausible hypotheses that miss the intended finding rather than generic or hallucinated text.

Table 5: Failure modes on the hard subset of RadDiffBench.

Failure mode and mechanism	Share of errors
Confounders: salient artifacts such as chest tubes distract from subtle pathology.	40%
Focal-to-diffuse: small lesions such as nodules are missed in favor of global patterns.	30%
Sign vs. cause: downstream signs such as an effusion are found, but their cause is missed.	20%
Other: miscellaneous errors.	10%

D.2 Qualitative Analysis

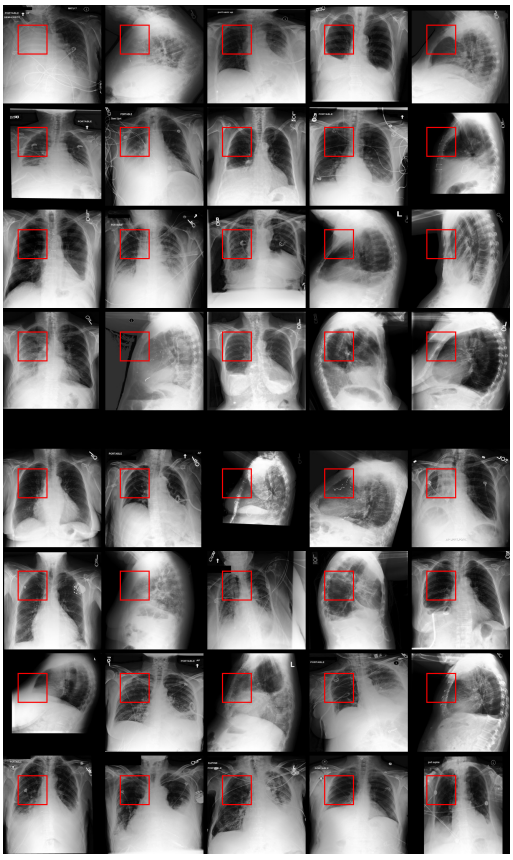
Iterative refinement. We illustrate how RadDiff converges toward stable, well-supported differences on the hard pair $\mathcal{R}_A$: *normal heart size* vs. $\mathcal{R}_B$: *cardiac silhouette remains enlarged*. In the

first iteration, the top candidates span several anatomical areas (Table 6, which extends Table 3 to the top five). Conditioned on these candidates and their scores, the refinement iteration converges on the ground truth: "*normal heart size and mediastinal/hilar contours in Group A*" and "*more consistent normal findings across Group A compared to Group B*."

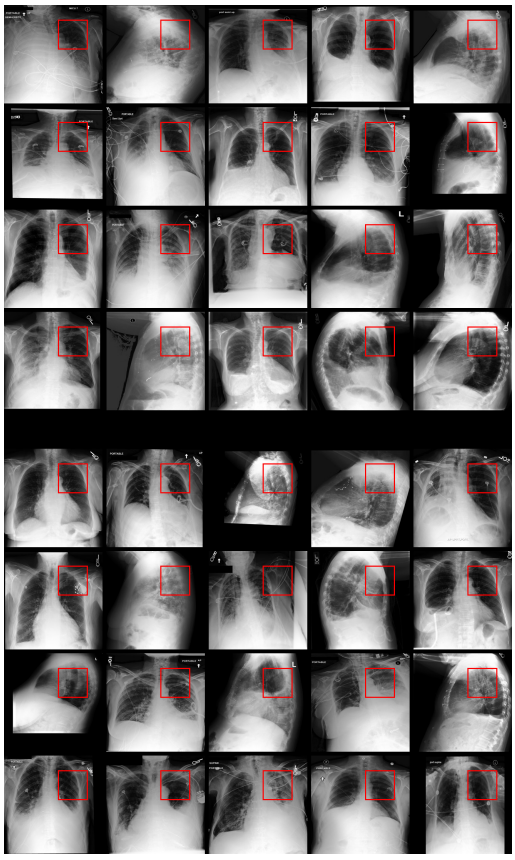
Table 6: Top candidates in the first iteration for *normal heart size* vs. *cardiac silhouette remains enlarged*. The candidates cover several areas; refinement then concentrates on normal heart size.

Rank	Predicted difference (first iteration)	$s(c)$
1	More cases with normal pulmonary vasculature in Group A	0.846
2	More cases showing normal osseous structures	0.844
3	More instances of normal heart size and mediastinal/hilar contours	0.830
4	More instances of normal heart size and mediastinal contours in Group A	0.823
5	More instances of normal or unchanged cardiomediastinal silhouette	0.815

Visual search. Figure 19 shows how visual search refines its focus on the pair *moderate right pleural effusion* vs. *no pleural effusion*. Across iterations, the selected regions concentrate on the right lower lung, where the effusion is visible, mirroring a radiologist's iterative inspection.



(a) RADDIFF progressively zooms into the right lower lung.



(b) RADDIFF scans alternative regions and finds no effusion-related signal.

Figure 19: Visual search on *moderate right pleural effusion* vs. *no pleural effusion*. Each panel is the exact 8×5 grid (40 images) shown to the proposer: the top half shows $\mathcal{R}_A$ and the bottom half $\mathcal{R}_B$; boxes mark the selected regions.

D.3 Case Study: From Candidates to Discovery

We trace RadDiff end to end on the medium pair $\mathcal{R}_A$: *dense right upper-lobe airspace consolidation* vs. $\mathcal{R}_B$: *no airspace consolidation*.

First iteration. Table 7 lists the top-ranked candidates and their verification scores.

Table 7: Top-5 candidates and verification scores for *dense right upper-lobe airspace consolidation* vs. *no airspace consolidation*.

Rank	Predicted difference (first iteration)	$s(c)$
1	More extensive bilateral parenchymal opacities	0.786
2	More widespread pulmonary opacities indicating multifocal pneumonia/edema	0.766
3	More reports of extensive bilateral pulmonary opacities	0.749
4	More bilateral pulmonary opacities present	0.702
5	Presence of large pleural effusions in Group A	0.680

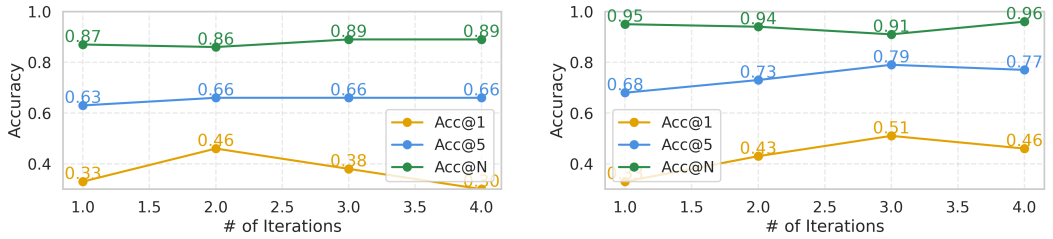
Visual search. For each of these candidates, visual search selects one region and zooms into it across all sampled studies (Figure 20).

After refinement. Conditioned on the verified candidates and the zoomed regions, RadDiff produces more anatomically specific and discriminative hypotheses:

- Distribution and extent of lung parenchymal abnormalities favoring large, bilateral consolidations in Group A.
- Less extensive bilateral opacities and absence of large pleural effusions in Group B.
- More extensive bilateral parenchymal opacities in Group A.
- More diffuse pulmonary edema pattern with diffuse bilateral opacities in Group A.
- Large pleural effusions with associated atelectasis predominantly in Group A.

D.4 Additional Ablations

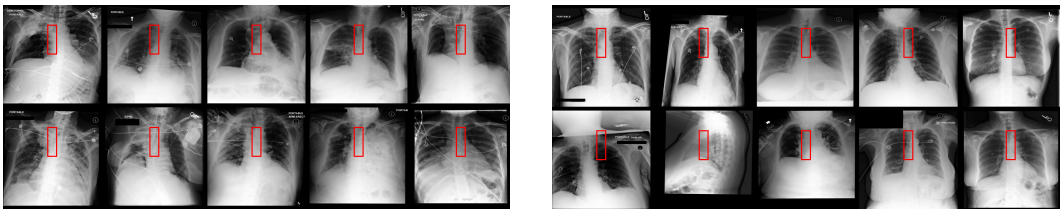
Figure 21 examines iterative refinement in two settings: (a) with generated captions and top-10 feedback, and (b) with summaries of the ground-truth reports. With generated captions, performance improves from the first to the second iteration on all metrics and then plateaus or declines slightly: early iterations contribute new refinements, whereas later ones mostly recycle overlapping hypotheses. Accordingly, Table 2 reports top-10 feedback after two iterations and top-5 feedback after three. With ground-truth reports, accuracy keeps increasing through the third iteration before stabilizing.



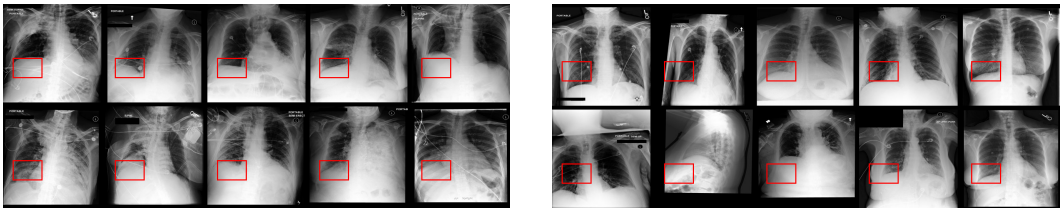
(a) Generated captions (top-10 feedback).

(b) Ground-truth reports.

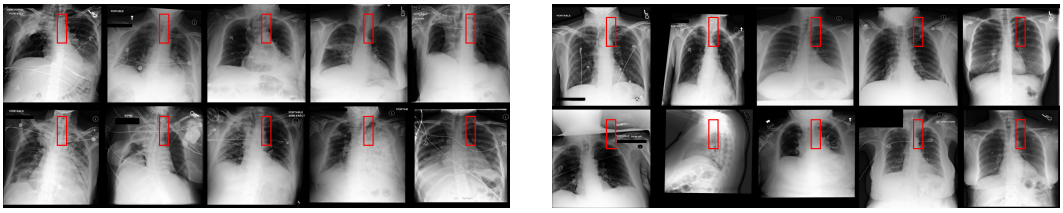
Figure 21: Accuracy across refinement iterations. In both settings, accuracy plateaus after the second or third iteration.



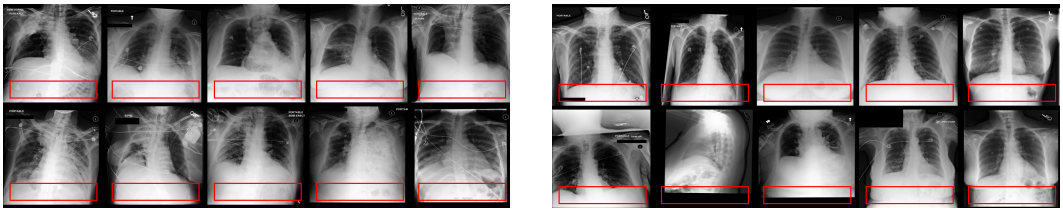
(a) Cropped by Top-1 candidate difference: *More extensive bilateral parenchymal opacities*



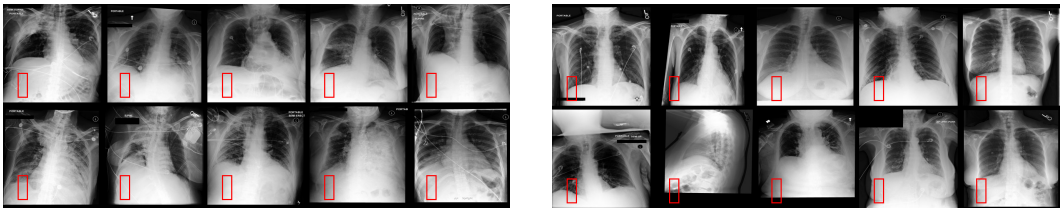
(b) Cropped by Top-2 candidate difference: *More widespread pulmonary opacities indicating multifocal pneumonia / edema*



(c) Cropped by Top-3 candidate difference: *More reports of extensive bilateral pulmonary opacities*



(d) Cropped by Top-4 candidate difference: *More bilateral pulmonary opacities present*



(e) Cropped by Top-5 candidate difference: *Presence of large pleural effusions in Group A*

Figure 20: Regions selected by visual search for each of the top-5 candidate differences. Left: $\mathcal{R}_A$; right: $\mathcal{R}_B$.

E Case Study Details

E.1 Patient Selection

Table 8 lists the ICD-9/10 codes used to select pneumonia and COVID-19 cases for the case studies in Section 6.

Table 8: Disease categories and ICD codes used for patient selection.

Disease	ICD codes	Description
Pneumonia	480–486; J13–J18, J851	All types
Heart failure	428*; I50*	Congestive heart failure
COPD	490–496; J40–J44	Chronic lung disease
Respiratory failure	518*; J96*	Acute/chronic
Sepsis	99591–99592; A41*, R652*	Septic states
ARDS	51882; J80	Acute respiratory distress

E.2 Pneumonia Mortality across Time Horizons

Table 9 quantifies how the hypotheses RadDiff generates for pneumonia non-survivors change with the mortality horizon, by grouping each comparison's hypotheses into clinical categories.

Table 9: Pneumonia non-survivors across mortality horizons. Each cell counts the hypotheses in a category among the n hypotheses generated after deduplication for in-hospital, 30-day, and 90-day mortality, each compared with patients who survived at least one year. As the horizon lengthens, acute markers decrease and chronic-failure signatures increase.

Category	Clinical feature	In-hospital (n=24)	30-day (n=19)	90-day (n=16)
Acute decompensation	Acute airway support	4/24	0/19	0/16
Unresolved injury	Diffuse parenchymal opacity	11/24	7/19	5/16
	Cardiomegaly	0/24	3/19	1/16
Chronic failure	Large pleural effusions	4/24	6/19	9/16
	Chronic drainage devices	5/24	3/19	4/16

E.3 Extended Results

We list the full ranked hypotheses RadDiff generates for three additional comparisons.

Pneumonia: 90-day death vs. in-hospital death. The main text contrasts in-hospital deaths with long-term survivors, which is dominated by devices and interventions. Comparing 90-day with in-hospital deaths instead surfaces parenchymal patterns (hyperinflation, atelectasis, effusion severity), a complementary view of survival-related imaging differences.

- Hyperinflation with diaphragmatic flattening significantly more common.
- Hyperinflated lungs with more prominent diaphragmatic flattening.
- Presence of hyperinflation and flattening of the hemidiaphragms.
- Smaller or absent pleural effusions.
- Less evidence of pulmonary fibrosis features.
- Normal cardiomediastinal silhouette without cardiomegaly.
- Enlarged mediastinal silhouette noted in some cases.
- Emphysema with low lung volumes and flattened diaphragms versus extensive bilateral opacities and pneumonia.
- Bibasilar atelectasis versus more localized or extensive atelectasis.
- Frequent bibasilar and bilateral atelectasis.
- Less frequent pulmonary edema or focal pneumonia.
- More normal pulmonary vasculature and mediastinal contours versus mild edema or cardiomegaly.
- Bilateral pulmonary hyperinflation more frequent.
- Focal consolidation and localized pneumothorax more frequent in some cases.
- Normal cardiomediastinal silhouette versus cardiomegaly with retrocardiac atelectasis and mild edema.
- Moderate left pleural effusion with adjacent atelectasis.
- Small pneumothorax more frequently noted.
- Small right apical pneumothorax versus small-to-moderate contralateral pneumothorax.
- Compressed or shifted mediastinum.
- Bilateral pleural effusions with bibasilar atelectasis more prevalent.
- Presence of right-sided central line with tip at cavoatrial junction.

- Interval removal of chest tubes without pneumothorax.
- More instances of small pneumothorax.
- Small right pleural effusion with overlying atelectasis versus moderate effusions with compressive atelectasis.
- Absence of bilateral parenchymal opacities and pneumonia versus extensive bilateral opacities.
- Widespread bilateral parenchymal opacities more frequent.
- Extensive bilateral parenchymal opacities indicating edema or atelectasis.
- Absence of widespread pulmonary opacities and consolidations in some cases.

Race: Asian vs. White (classifier-selected). The main text compares White vs. Asian subsets and finds that the Asian subset is characterized by normal findings and the White subset by abnormalities. The full list of hypotheses with the Asian subset as $\mathcal{R}_A$ is:

- Small right apical pneumothorax more frequently observed.
- More cases showing no focal consolidation or pneumothorax.
- More images with well-expanded lungs.
- More mentions of a large right upper-lobe mass with air bronchograms.
- Repeated absence of focal consolidation or pneumothorax.
- More normal overall findings with occasional complications.
- More normal mediastinal and hilar contours.
- Increased lung volumes, including low lung volumes.
- Normal lung volumes with some low-volume cases.
- More normal lung findings despite complications.
- More normal lung findings overall with limited complications.
- Less frequent emphysema with flattened hemidiaphragms.
- More normal heart size and vascular structures.
- Normal heart size and vasculature more common.
- More frequent large right pleural effusion with atelectasis.
- More low lung volumes or hyperinflation.
- More cases without pleural effusion or pneumothorax, except isolated ones.
- More complications such as pneumonia or loculated effusions.
- More low-volume lungs with bibasilar atelectasis.
- More mentions of interventions such as endotracheal tubes or catheters.
- Fewer cases with large hiatal hernia or major abnormalities.
- More small left pleural effusions with atelectasis.
- Higher frequency of small bilateral pleural effusions.
- Normal cardiomediastinal silhouette with vascular congestion and edema more frequently described.

Race: White vs. Black (classifier-selected). The differences follow the same qualitative pattern as the White–Asian comparison:

- Higher prevalence of pleural effusions and pneumothorax.
- Presence of endotracheal tube tips at or above the clavicles noted more often.
- More cases of small bilateral pleural effusions with atelectasis.
- More cases with pulmonary vascular congestion and interstitial edema.
- More frequent projection of tubes over the stomach or above the carina.
- Bilateral small pleural effusions with overlying atelectasis.
- Pleural effusions, including large unilateral or left-sided effusions.
- Tips of endotracheal and nasogastric tubes projecting over the stomach or above the carina.
- Large or unilateral left-sided pleural effusions more prevalent.
- Presence of a large hiatal hernia more frequently described.
- Reports of large hiatal hernia appearing in multiple cases.
- Small residual pneumothorax (left or right) noted more commonly.
- Hyperinflated lungs with flattening of the hemidiaphragms (emphysema).
- Presence of a hiatal hernia.
- Enlarged cardiac silhouette and signs of pulmonary edema more frequently observed.
- Greater incidence of hyperinflation with flattening of the hemidiaphragms.
- Consolidation in the left upper lobe more characteristic in some cases.
- Enlarged cardiomediastinal silhouette more commonly described.
- Lower lung volumes with accentuated cardiomediastinal silhouette.