Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Analysis of Clinical Finding Retrieval Difficulty in Chest X-ray Retrieval Models

Areesha farid

arishafarid29@gmail.com

Dow University of Health Sciences https://orcid.org/0009-0007-2133-8851

Hifsa Farooq

Dow University of Health Sciences https://orcid.org/0009-0001-9396-9472

Research Article

Keywords: chest X-ray, sentence retrieval, retrieval pool composition, finding-level evaluation, radiology report generation

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11003136/v3

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Comparative Analysis of Clinical Finding Retrieval Difficulty in Chest X-Ray Retrieval Models

Areesha Farid
Dow University of Health Sciences

Hifsa Farooq
Dow University of Health Sciences

Abstract

Sentence-level retrieval generates chest X-ray (CXR) reports by selecting clinical sentences from a predefined pool. Retrieval performance depends on the composition of this pool, yet aggregate metrics can conceal differences between clinical findings. This study examined whether finding-level retrieval difficulty reflects model limitations or retrieval pool composition. Three sentence-level models (a JoImTeRNet-based model developed by the authors, CXR-RePaiR and CXR-ReDonE) retrieved the top 10 sentences for 10,000 frontal CheXpert images. Sentences were drawn from an original pool of 129,906 MIMIC-CXR sentences, 30 balanced pools of approximately 6,200 sentences each (up to 450 positive sentences per finding, drawn with different random seeds) and 50 random pools of equal size. Sentences were labeled with the CheXbert labeler. Micro-averaged recall, precision and F1, finding-level recall at $K = 5$ and cross-model Spearman correlation were evaluated. Balancing raised micro-recall for every model at every K and finding-level recall for 13 of 14 findings, whereas No Finding recall fell by 0.119 to 0.187 (mean of 30 balanced pools). For 36 of 42 finding-model combinations the change exceeded that of random pools in at least half of the balanced pools, and at $K = 5$ balanced-pool recall exceeded the random-pool mean by 0.07 to 0.08. Precision decreased and F1 varied. Cross-model agreement across 13 findings declined from $\rho = 0.687$–0.808 (original pool) to $\rho = 0.153$–0.577 (balanced pools), while random pools gave $\rho = 0.712$–0.788; with only 13 findings, this decrease should be regarded as descriptive. Lung Lesion remained among the five lowest-recall findings for all three models under the original pool and in all 30 balanced pools. Retrieval pool composition influences finding-level recall and apparent difficulty in CXR sentence retrieval, and part of the difficulty observed under the full pool is pool-dependent. Evaluations should report pool composition and finding-level results alongside aggregate scores.

Keywords chest X-ray · sentence retrieval · retrieval pool composition · finding-level evaluation · radiology report generation

1 Introduction

Automated chest X-ray (CXR) report generation aims to reduce the workload associated with interpreting radiographs and producing diagnostic reports [1, 2]. Unlike generative approaches that produce text directly from images, retrieval-based methods select relevant clinical text from an existing corpus [3]. These methods can retrieve complete reports, report sections, or individual sentences to describe radiographic findings [5, 8].

Sentence-level retrieval allows systems to select individual clinical statements from a predefined retrieval pool, making retrieval pool composition an important consideration when evaluating retrieval quality.

Retrieval performance depends not only on a model's ability to identify relevant sentences but also on the composition of the retrieval pool. The available candidates determine the alternatives against which relevant sentences must be retrieved. Previous research has shown that enlarging retrieval pools with visually similar images can reduce recall across models [15]. Similarly, changes in evaluation-set composition can affect measured performance even when the model remains unchanged [16]. These findings highlight the importance of considering retrieval pool composition when interpreting retrieval performance. Furthermore, aggregate metrics can conceal differences in performance across individual clinical findings [5]. Consequently, overall retrieval scores alone may not reveal whether a finding is difficult to retrieve because of model limitations or because of the composition of the retrieval pool. CXR research has examined retrieval methods and differences in performance across clinical findings [5, 8]. Other studies have demonstrated that retrieval performance can vary with retrieval pool construction [15] and that evaluation-set composition can influence measured performance in medical imaging [16].

Existing research has investigated CXR retrieval architectures, finding-level performance variation, and the sensitivity of retrieval or medical image evaluation to retrieval-pool and test-set composition. However, the relationship between retrieval pool composition and finding-level retrieval difficulty in CXR sentence retrieval remains insufficiently characterized. In particular, it is unclear whether findings that are difficult to retrieve under one retrieval pool composition remain difficult under another, or whether their relative difficulty changes with the available candidate sentences. This distinction matters because low retrieval performance may reflect model limitations, the composition of the retrieval pool, or an interaction between them. The present study addresses this question by comparing three sentence-level CXR retrieval models under different retrieval pool compositions and examining both aggregate performance and finding-level retrieval difficulty.

2 Related work

Chest X-ray image-to-text retrieval treats report generation as a search problem, in which relevant clinical text is retrieved from an existing corpus instead of being generated freely [8]. Early approaches selected complete reports or report sections. X-REM, for example, retrieves report sections using a multimodal image-text matching score and applies a natural-language-inference filter to remove redundant or contradictory text [8]. CXR-RePaiR moved retrieval to the sentence level, selecting sentences from a fixed retrieval pool with a pre-trained contrastive language-image model [5]. JoImTeRNet [22] learns joint image-text representations that were assessed through cross-modality retrieval, and CXR-ReDonE [21] is a retrieval-based report generation model. Later work has explored ways to improve what is retrieved. DrLS combines diversity-controlled retrieval with a medical lexicon and a sentence-level semantic loss [9], and retrieved text has also been used as context for GPT-based report generation [10]. Beyond report generation, image-to-text and text-to-image retrieval are also used to assess chest X-ray vision-language representations, with recall at K measuring how well images and text are matched [11].

Performance is commonly reported with aggregate metrics computed over a whole test set or across all target classes at once. CXR-RePaiR, for instance, reports a single macro-averaged F1 across 14 radiological categories, together with semb and BLEU2, to compare against generative baselines [5]. BenchX standardizes evaluation across nine medical vision-language pre-training methods by reporting one averaged score per dataset and task and condensing these into an overall ranking [14]. Some studies also break results down by individual finding, which shows variation that aggregate scores do not. CXR-RePaiR's per-finding analysis

reports F1 of around 0.6 for pleural effusion against about 0.15 for consolidation under the same setup [5]. KAD reports a comparable spread on PadChest, reaching an AUC of at least 0.900 on 31 unseen classes but only 0.700 on 111 of 177 unseen classes overall [12]. ProtoCLIP shows the same pattern at smaller scale, with an AUC of 0.96 for pleural effusion against 0.82 to 0.84 for atelectasis and lung opacity [13].

Results in cross-modal retrieval can vary with how the retrieval pool is built. State-of-the-art models reach near-perfect recall on standard MSCOCO and Flickr30K splits, a result that largely reflects coarse retrieval pools. Recall dropped substantially across all tested architectures when the pool was enlarged with visually similar images or when captions were refined with finer detail [15]. Comparable sensitivity has been reported for evaluation sets in medical imaging. Precision, recall and F1 for pediatric wrist fracture detection were significantly lower on a test set balanced for case difficulty than on a randomly sampled one, despite no change to model training [16]. In chest radiography, the reference used to define pathology labels, image-derived or report-derived, changes both absolute performance and the relative ranking of supervised and vision-language models, with the degree of disagreement differing by pathology [17].

Several retrieval studies have compared multiple models under shared conditions to look beyond aggregate scores. A comparison of four visual-semantic embedding networks on Flickr30K found that the two best models shared 16 of their 25 worst-performing image classes, and that they disagreed on whether a relevant description was retrieved for a substantial share of queries [18]. ABNIRML probed lexical, dense and transformer-based ranking models with diagnostic tests, including several random seeds of the same model [19], and a reproducibility study evaluated four vision-language models under shared textual perturbations across datasets of different granularity [20].

3 Materials and Methods

3.1 Retrieval Models

3.1.1 Our model

We developed a sentence-level image-to-text retrieval model based on the JoImTeRNet framework proposed by Ji et al., which learns a shared embedding space between chest X-rays and radiology reports through global image-sentence matching and local region-word matching. Our implementation retained the global sentence-level matching (TIM) and word-region alignment (WRA) branches, while omitting the original bigram/trigram phrase-region alignment branch. Masked language modeling (MLM) was added as an auxiliary training objective.

The model was trained on MIMIC-CXR using 30,432 radiology reports paired with frontal chest X-rays. The Findings sections were segmented into sentences using spaCy's `en_core_web_sm` tokenizer, and sentences with fewer than three words were removed, resulting in 221,070 image-sentence pairs. To reduce the influence of frequently repeated normal and templated sentences, each unique sentence was limited to three image-sentence instances. Excess instances were randomly sampled using seed 42, leaving 148,032 pairs after deduplication, representing a 33.0% reduction. Data were split at the image level into training, validation, and test sets using a 90/5/5 ratio, resulting in 133,158, 7,452, and 7,422 image-sentence pairs, respectively. The training set contained 26,388 unique images. Since an image could have multiple associated sentences, a custom sampler selected one sentence per image per epoch to ensure that each image contributed at most one sentence to a training batch.

Chest X-rays were converted to grayscale, resized to 256×256 pixels, and normalized using a mean and

standard deviation of 0.5. Training images were augmented with random horizontal flipping and rotations of $\pm 10^{\circ}$. The image encoder was an ImageNet-pretrained ResNet-50, whose first convolutional layer was adapted for single-channel input by summing the pretrained RGB weights. Region features were extracted from the third residual block at 16×16 spatial resolution, processed with soft attention, and projected to 512 dimensions, producing 256 region vectors. Global features were obtained from the fourth residual block using global average pooling and a 512-dimensional projection. Sentences were tokenized using the BERT-base-uncased WordPiece tokenizer and padded or truncated to 128 tokens. The text encoder consisted of three Transformer layers with eight attention heads and a hidden size of 512. It was initialized from a randomly configured `BertConfig`, meaning that pretrained BERT weights were not used. The [CLS] representation was projected to form the sentence embedding, while token-level representations were projected to 512-dimensional word embeddings for region-word matching.

Image-text alignment was optimized at both sentence and word levels. Sentence-level matching used bidirectional cross-entropy losses with cosine similarities scaled by $\gamma_3 = 10$. Word-region alignment followed the JoImTeRNet WRA formulation with attention and similarity temperatures of $\gamma_1 = 4$ and $\gamma_2 = 5$. Bidirectional triplet losses with a margin of 0.3 were applied at both levels. Hard negatives were mined within each batch using the most similar non-matching sample based on sentence-level cosine similarity. An auxiliary MLM objective was trained jointly with retrieval. Fifteen percent of non-special tokens were selected for masking; 90% were replaced by [MASK], approximately 10% by random vocabulary tokens, and the remainder were left unchanged. The MLM head predicted the original tokens using cross-entropy loss. Training updates were sampled with probabilities of 0.7 for image-text matching and 0.3 for MLM. For image-text matching, sentence- and word-level matching losses were weighted by 2.0, while triplet losses were weighted by 1.0. The image and text encoders were optimized separately using AdamW. Learning rates were 1×10^{-5} for the third and fourth ResNet-50 blocks and 1×10^{-4} for newly initialized image layers and the text encoder. Both optimizers used a weight decay of 1×10^{-4} and cosine annealing over 30 epochs with a minimum learning rate of 1×10^{-7}. Gradient norms were clipped at 1.5. Training used a batch size of 24, with 16 samples for validation, on a single GPU in Kaggle. After each epoch, validation image-to-text Recall@1 was computed, and the checkpoint with the highest Recall@1 was retained.

3.1.2 CXR-RePaiR

CXR-RePaiR (Contrastive X-ray-Report Pair Retrieval) is a retrieval-based chest X-ray report generation method proposed by Endo et al. Rather than generating reports autoregressively, it retrieves text from a predefined corpus based on the similarity between the input X-ray and candidate text. The method uses a CLIP-based image-text encoder to map images and text into a shared embedding space, where similarity is calculated between the corresponding embeddings. For our experiments, we used the publicly released CXR-RePaiR checkpoint `clip-imp-pretrained_128_6_after_4.pt` from the official repository. This checkpoint corresponds to the pretrained and radiology-adapted configuration reported by Endo et al. We used the model without any additional training or fine-tuning, keeping both the image and text encoders frozen and operating them in evaluation mode.

To ensure consistency with our sentence-level retrieval setting, we used CXR-RePaiR to retrieve individual sentences rather than complete reports. The retrieval corpus was constructed from the findings sections of the MIMIC-CXR reports used for training our model. Sentences were extracted using the same preprocessing procedure described in Section 3.1.1, including spaCy-based sentence segmentation and a minimum length of three words, resulting in 129,906 unique sentences. The released checkpoint's original CLIP preprocessing pipeline was retained. Images were resized and center-cropped to 224×224 pixels, converted from grayscale

to three-channel RGB, and normalized using the standard CLIP mean and standard deviation. Text was tokenized using the CLIP tokenizer with a context length of 77 tokens. All sentences in the retrieval corpus were encoded once using the frozen CXR-RePaiR text encoder and L2-normalized. For each input image, the corresponding image embedding was generated and similarly normalized. Sentence-image similarity was then computed using the dot product between the normalized embeddings, and the highest-scoring sentences were selected as the retrieved outputs.

3.1.3 CXR-ReDonE

CXR-ReDonE, proposed by Ramesh, Chi, and Rajpurkar (2022), builds on the CXR-RePaiR retrieval framework by replacing its CLIP-based model with an ALBEF-based vision-language model. The model uses a Vision Transformer as the image encoder to obtain patch-level image representations and a BERT-based transformer as the text encoder to obtain token-level text representations. Global image and text embeddings are obtained from the respective [CLS] representations. CXR-ReDonE also includes a multimodal encoder composed of additional transformer layers with cross-attention, which combines image and text representations for image-text matching. For retrieval, however, the model uses the global image and text embeddings directly, with similarity calculated using their dot product. The highest-scoring report or sentence is selected as the retrieval output, with top-k sentences retrieved when multiple outputs are required. The training objective follows the ALBEF framework and combines image-text contrastive learning, image-text matching, and masked language modeling. The contrastive objective aligns global image and text representations before multimodal fusion, while the image-text matching objective uses the cross-attention encoder together with hard-negative mining. Momentum distillation is also used to provide soft pseudo-targets during contrastive training.

A key difference between CXR-ReDonE and CXR-RePaiR lies in the training data. CXR-ReDonE is trained on MIMIC-PRO, an adapted version of MIMIC-CXR in which references to previous studies, such as "unchanged," "stable," and "compared to prior," are removed at the token level using the GILBERT model. The model is trained using image-impression pairs from this modified corpus. The reported training procedure consists of 60 epochs with 100 steps per epoch using four Quadro RTX-8000 GPUs. For our experiments, we used the pretrained `checkpoint_59.pth` released through the official CXR-ReDonE repository. This checkpoint corresponds to the model trained on MIMIC-PRO with prior-study references removed. It was loaded into the ALBEF retrieval architecture provided in the repository through `models.model_retrieval.ALBEF`, together with the supplied `Retrieval_flickr.yaml` and BERT configuration files. The visual encoder position embeddings were interpolated to match the spatial resolution stored in the checkpoint. The released checkpoint was used without further training or fine-tuning. Both the image and text encoders remained frozen throughout inference, with gradients disabled and the model operating in evaluation mode.

Images were first resized so that the longer side measured 320 pixels (Lanczos interpolation) and centered on a 320×320 canvas to preserve the aspect ratio. They were then resized to 256×256 pixels using bicubic interpolation and converted from grayscale to three-channel input. The checkpoint's original pixel normalization was retained, using per-channel mean and standard deviation values of approximately 101.49 and 83.44 applied directly to the raw 0–255 pixel values. This preprocessing follows the HDF5-based pipeline used during the model's original training rather than applying conventional normalization after rescaling pixel values to the 0–1 range. Text was tokenized using the `bert-base-uncased` tokenizer and padded or truncated to a maximum length of 64 tokens. For sentence-level retrieval, each sentence in our retrieval corpus was encoded once using the text encoder's [CLS] representation and its corresponding text projection head. The resulting sentence embeddings were used as the retrieval index. Each evaluation image was passed

through the visual encoder and projection head to obtain its global [CLS] image representation. Image and sentence embeddings were then compared using cosine similarity, and sentences were ranked according to their similarity to the input image.

3.2 Evaluation Dataset

We used the CheXpert dataset as the evaluation dataset for all retrieval models. Only frontal-view chest radiographs were included, together with their corresponding structured clinical annotations. A random sample of 10,000 frontal-view images was drawn without replacement from the CheXpert training set using a fixed random seed of 42. The resulting dataset retained the class distribution of the original training set. Ground-truth annotations were obtained directly from the structured CheXpert labels, which represent each observation as positive, negative, uncertain, or unmentioned. Unmentioned findings were treated as negative. Uncertain annotations were retained as uncertain, were not counted as positive, and were excluded from the micro-averaged precision, recall and F1 and from finding-level recall.

3.3 Sentence Pool Construction

The retrieval sentence pool was constructed from the findings sections of the MIMIC-CXR reports used for training our model. Sentences were extracted using the same spaCy-based procedure described in Section 3.1.1, with a minimum length of three words and no deduplication cap applied at this stage. This resulted in 129,906 unique retrieval pool sentences. Each sentence was labeled using the CheXbert labeler based on the 14 CheXpert clinical findings. These sentence-level labels were used to characterize the distribution of findings within the pool and to construct the balanced and random pools. Because a sentence can carry more than one finding, the percentages of sentences per finding sum to more than 100%. For the balanced pool, the number of positive sentences for each finding was capped at 450. All 14 findings had more than 450 positive sentences in the original pool, the smallest being Pleural Other with 530, so 450 sentences were randomly selected for every finding (first draw, seed 42). Since individual sentences could carry multiple findings, the sampled sentences were combined across findings and duplicate sentences were removed. This produced a balanced pool of 6,208 unique sentences, with between 454 (Pleural Other) and 1,339 (Lung Opacity) sentences per finding. Because the cap equalizes the number of sampled sentences per finding but not the number of sentences carrying each finding, this pool is capped rather than strictly balanced, and the term balanced is used for brevity. To assess sampling variability, the construction was repeated with 30 further random seeds, and results for the balanced pool are reported as the mean ($\pm$ SD where shown) across these 30 pools. The balanced pool differs from the original pool in both class balance and total size (129,906 vs. 6,208 sentences). To separate these two factors, random pools of the same size as the balanced pool were also constructed (Section 3.4).

3.4 Random Size-Matched Pools

To control for pool size, 50 random pools were drawn, each containing 6,208 sentences, the size of the balanced pool. Sentences were sampled without replacement from the full 129,906-sentence pool without regard to their labels (random number generator seed 123), so that each random pool retains the finding distribution of the original pool in expectation. Results for random pools are reported as the mean and standard deviation across the 50 draws.

3.5 Evaluation Metrics

For each of the 10,000 evaluation images, the top 10 sentences were retrieved from the pool under evaluation (the original pool, the balanced pool, or one of the random pools) using cosine similarity between the image embedding and the pre-computed sentence embeddings of each model. Image embeddings were computed once per model and reused for all pools, so changing the pool changed only the set of sentences searched. Pool sentences were tokenized to a maximum of 64 tokens for our model and CXR-ReDonE and 77 tokens for CXR-RePaiR. Retrieved sentences were not labeled again. Each retrieved sentence took the CheXbert labels assigned to it in the pool.

For a given K, a finding was predicted present for an image if at least one of the top K retrieved sentences was positive for that finding, and predictions were compared with the CheXpert labels of the image. Precision, recall and F1 were micro-averaged across all image-finding pairs, excluding pairs with an uncertain ground-truth label. Finding-level recall was calculated as the proportion of images with a positive label for the finding in which the finding was predicted. Finding-level analyses used $K = 5$.

The change in finding-level recall between pools (Δ = balanced $-$ original) was compared with the change obtained for each of the 50 random pools (random $-$ original). A balanced-pool change was considered to go beyond the effect of pool size if it fell outside the 2.5th to 97.5th percentile range of the random changes in at least half of the 30 balanced pools. Agreement between models in finding-level difficulty was measured with the Spearman rank correlation of finding-level recall at $K = 5$, calculated over the 13 findings excluding No Finding. For the original pool, confidence intervals were obtained from 3,000 bootstrap resamples of the evaluation images. For the balanced and random pools the correlation was calculated for each pool; balanced values are given as the mean with the 2.5th to 97.5th percentile across the 30 pools, and random values as mean $\pm$ SD. To test whether the decrease in agreement was robust to the choice of findings, the difference in ρ between the original and balanced pools ($\Delta\rho$) was bootstrapped over the 13 findings (5,000 resamples). A finding was considered hard for a model if it was among the five lowest-recall findings of the 13 findings other than No Finding at $K = 5$, and hard for all three models if this held for each model.

4 Results

Overall retrieval performance varied across retrieval pool conditions, with different patterns in recall, precision, and F1 across the three models (Table 1). Recall and F1 generally increased with K, whereas precision decreased. Under the original retrieval pool, CXR-RePaiR achieved the highest recall, precision, and F1 at each K. At $K = 5$, its recall, precision, and F1 were 0.5299, 0.4100, and 0.4623, respectively.

The balanced retrieval pools (mean of 30 draws) produced higher recall than the original pool for all three models at every K, although precision decreased. Changes in F1 were smaller and varied by model and K. To distinguish the effect of pool composition from the effect of pool size, we compared the balanced pool with 50 random pools of the same size. At $K = 5$, recall in the balanced pool exceeded the random-pool mean by 0.0817 for our model, 0.0806 for CXR-RePaiR, and 0.0715 for CXR-ReDonE. The standard deviations across random pools ranged from 0.0087 to 0.0130 for recall at this K, and across balanced pools from 0.0090 to 0.0174. Thus, the balanced-pool results differed from those obtained using randomly sampled pools of the same size.

The original retrieval pool contained 129,906 sentences with an uneven distribution across 14 clinical findings (Figure 1). No Finding accounted for 39.0% of sentences, whereas Pleural Other represented only 0.4%. A representative balanced pool (seed 42) contained 6,208 sentences, increasing the representation of Pleural

Table 1: Overall retrieval performance (micro-averaged recall, precision, F1) at $K = 1$, 5, and 10 under the original, balanced, and random retrieval pools. Balanced rows give the mean ± standard deviation (SD) across 30 balanced pools, and random rows across 50 random pools of the balanced pool's size.

		$K = 1$			$K = 5$			$K = 10$		
Model	Pool	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Our model	Original	0.1858	0.3720	0.2478	0.3937	0.3594	0.3758	0.4932	0.3445	0.4056
	Balanced	0.2186 ±0.0111	0.3564 ±0.0116	0.2710 ±0.0117	0.4882 ±0.0169	0.3051 ±0.0074	0.3755 ±0.0099	0.6188 ±0.0183	0.2825 ±0.0057	0.3879 ±0.0082
	Random	0.1865 ±0.0077	0.3744 ±0.0095	0.2490 ±0.0085	0.4065 ±0.0095	0.3601 ±0.0057	0.3819 ±0.0060	0.5167 ±0.0128	0.3448 ±0.0056	0.4135 ±0.0061
CXR-RePaiR	Original	0.2919	0.5000	0.3686	0.5299	0.4100	0.4623	0.6304	0.3742	0.4696
	Balanced	0.3128 ±0.0155	0.4202 ±0.0108	0.3585 ±0.0132	0.5855 ±0.0174	0.3467 ±0.0062	0.4355 ±0.0088	0.6964 ±0.0160	0.3152 ±0.0050	0.4339 ±0.0070
	Random	0.2574 ±0.0114	0.4814 ±0.0112	0.3353 ±0.0101	0.5049 ±0.0130	0.4046 ±0.0065	0.4491 ±0.0062	0.6146 ±0.0130	0.3712 ±0.0061	0.4628 ±0.0057
CXR-ReDonE	Original	0.1868	0.3997	0.2547	0.3835	0.3796	0.3815	0.4786	0.3675	0.4158
	Balanced	0.2237 ±0.0075	0.3888 ±0.0114	0.2840 ±0.0088	0.4797 ±0.0090	0.3299 ±0.0050	0.3909 ±0.0057	0.6100 ±0.0087	0.3017 ±0.0037	0.4037 ±0.0043
	Random	0.1887 ±0.0072	0.3945 ±0.0118	0.2553 ±0.0088	0.4082 ±0.0087	0.3779 ±0.0059	0.3925 ±0.0059	0.5213 ±0.0093	0.3627 ±0.0051	0.4277 ±0.0047

Other from 0.4% to 7.3% and Lung Lesion from 1.3% to 7.9%, while reducing No Finding from 39.0% to 14.0%. Random pools of the same size retained the original distribution in expectation, enabling comparison of pool composition effects independently of pool size.

Finding-level recall at $K = 5$ differed substantially across clinical findings and retrieval pool conditions (Table 2). Under the original retrieval pool, No Finding had the highest recall for all three models, whereas Lung Lesion, Pleural Other, and Fracture were consistently among the lowest-recall findings. Across the models, recall for these three findings ranged from 0.059 to 0.156.

Under the balanced retrieval pools, mean recall increased for 13 of the 14 findings across all three models. The only exception was No Finding, for which recall decreased by 0.119 to 0.187. The largest gains varied by model and finding. For our model, recall increased most for Pneumothorax (+0.398), Enlarged Cardiomediastinum (+0.337), and Fracture (+0.322). For CXR-RePaiR, the largest gains occurred for Consolidation (+0.216), Pneumonia (+0.215), and Pleural Other (+0.209). For CXR-ReDonE, the largest gains occurred for Pleural Other (+0.211), Enlarged Cardiomediastinum (+0.208), and Pneumothorax (+0.206). Support Devices showed comparatively small gains across all three models.

To separate these changes from the effect of pool size, the change in recall under the balanced pool was compared with the change observed under each of the 50 random pools (Table 3). The random pools produced mean changes between -0.099 and +0.085 across the 42 finding-model combinations. The balanced-pool change fell outside the random 95% range in at least half of the 30 balanced pools in 36 of the 42 combinations (12 of 14 findings for each model; 77% to 100% of pools). The exceptions were Lung Opacity (outside the range in 27% to 47% of pools) and Support Devices (3% to 17%) in all three models. In most cases the balanced-pool change was positive while the random-pool change was close to zero, as for Lung Lesion (+0.178, +0.163, and +0.165 against random means of +0.004, +0.007, and -0.009). Two outside-range cases differ in kind. For No Finding the balanced-pool change was negative while the random-pool change was positive, and for Pleural Effusion in CXR-RePaiR the balanced-pool change (+0.040) fell outside the random range because the random pools reduced recall (mean -0.071).

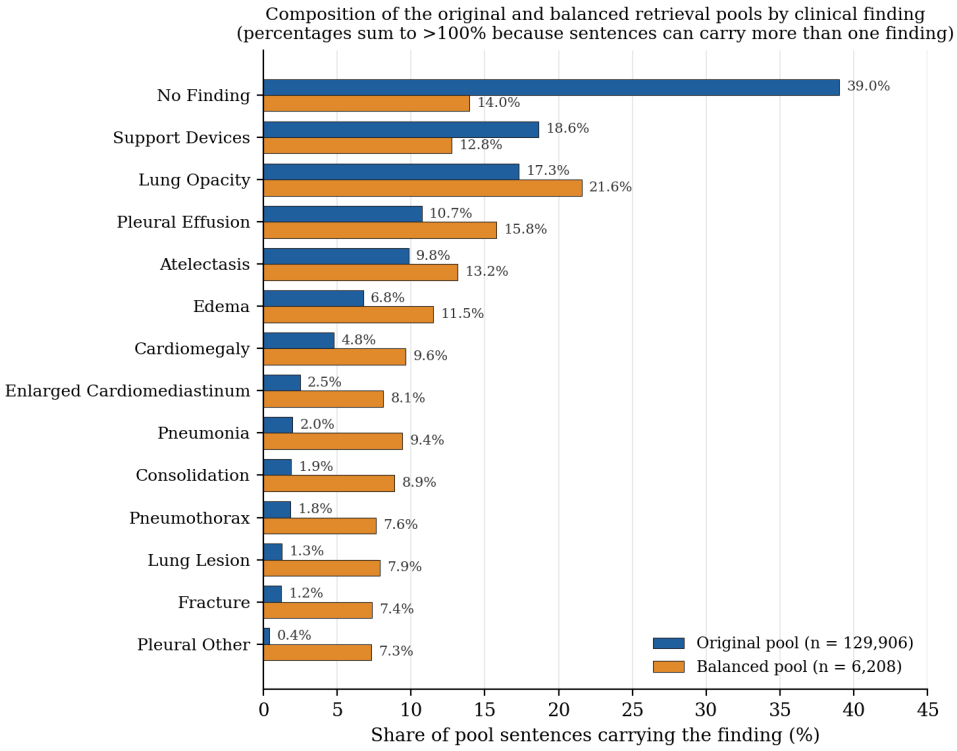


Figure 1: Composition of the original ($n = 129{,}906$ sentences) and balanced (seed 42 draw; $n = 6{,}208$ sentences) retrieval pools by clinical finding, shown as the share of pool sentences carrying each finding. Percentages sum to more than 100% because a sentence can carry more than one finding. Random pools of the balanced pool's size retain the original proportions in expectation.

Agreement between models in the relative difficulty of findings is shown in Table 4. Under the original pool, finding-level recall was strongly correlated for all three model pairs ($\rho = 0.687$–0.808 over 13 findings). Random pools gave similar agreement (mean $\rho = 0.712$–0.788). Under the balanced pools, agreement was lower for all three pairs (mean $\rho = 0.528$ for our model with CXR-RePaiR, 0.577 for CXR-RePaiR with CXR-ReDonE, and 0.153 for our model with CXR-ReDonE), and each balanced-pool mean lay 2.5 to 6.0 random-pool SDs below the corresponding random-pool mean. However, with only 13 findings the decrease was not statistically established: a bootstrap over findings gave $\Delta\rho = +0.213$ (95% CI -0.194 to $+0.779$) for our model with CXR-RePaiR and $+0.597$ (95% CI -0.011 to $+1.280$) for our model with CXR-ReDonE.

Findings classified as hard (among the five lowest-recall findings of the 13 findings other than No Finding at $K = 5$) differed between pool conditions. Under the original pool, Lung Lesion, Pleural Other, and Fracture were hard for all three models; the remaining two positions were Pneumonia and Consolidation for our model and CXR-RePaiR, and Pneumothorax and Enlarged Cardiomediastinum for CXR-ReDonE. Under the balanced pools, Lung Lesion was the only finding that was hard for all three models in all 30 pools (mean recall 0.250 to 0.272). Pleural Other and Fracture were among the lowest five for CXR-RePaiR (30 and 30 of 30 pools) and CXR-ReDonE (27 and 30 of 30 pools) but not consistently for our model (18 and 0 of 30 pools), in which Consolidation (29 of 30) and Edema (18 of 30) were more often among the lowest five. Pneumonia was hard for our model and CXR-RePaiR in 30 and 29 of 30 pools, and for CXR-ReDonE in 13 of 30.

Table 2: Finding-level recall at $K = 5$ under the original (Orig) and balanced (Bal, mean of 30 pools) retrieval pools for all three models. Δ = Bal $-$ Orig.

	Our model			CXR-RePaiR			CXR-ReDonE		
Finding	Orig	Bal	Δ	Orig	Bal	Δ	Orig	Bal	Δ
No Finding	0.956	0.770	−0.187	0.664	0.545	−0.119	0.876	0.746	−0.130
Enlarged Cardiomediastinum	0.197	0.535	+0.337	0.289	0.429	+0.140	0.113	0.321	+0.208
Cardiomegaly	0.243	0.405	+0.162	0.475	0.509	+0.033	0.273	0.432	+0.159
Lung Opacity	0.337	0.380	+0.043	0.520	0.561	+0.040	0.319	0.364	+0.045
Lung Lesion	0.083	0.261	+0.178	0.087	0.250	+0.163	0.107	0.272	+0.165
Edema	0.247	0.360	+0.114	0.415	0.513	+0.098	0.466	0.603	+0.138
Consolidation	0.076	0.262	+0.185	0.256	0.472	+0.216	0.180	0.383	+0.203
Pneumonia	0.044	0.187	+0.143	0.173	0.388	+0.215	0.149	0.325	+0.176
Atelectasis	0.225	0.379	+0.154	0.455	0.553	+0.099	0.226	0.337	+0.111
Pneumothorax	0.173	0.572	+0.398	0.596	0.762	+0.166	0.091	0.297	+0.206
Pleural Effusion	0.378	0.501	+0.122	0.650	0.691	+0.040	0.375	0.523	+0.147
Pleural Other	0.088	0.363	+0.275	0.105	0.314	+0.209	0.079	0.289	+0.211
Fracture	0.156	0.479	+0.322	0.097	0.258	+0.160	0.059	0.222	+0.163
Support Devices	0.666	0.681	+0.015	0.647	0.657	+0.010	0.552	0.599	+0.047

5 Discussion

This study examined whether finding-level difficulty in sentence-level chest X-ray (CXR) retrieval is attributable to the retrieval model or to the pool from which sentences are retrieved. Three main results emerged. First, balancing the retrieval pool increased mean recall for 13 of the 14 findings in all three models, and for 36 of 42 finding-model combinations the change fell outside the range produced by 50 size-matched random pools in at least half of 30 balanced pools. Second, balancing altered which findings were difficult to retrieve and reduced cross-model agreement. Spearman correlations across 13 findings decreased from 0.687–0.808 under the original pool to between 0.153 and 0.577 under the balanced pools, whereas random pools produced agreement similar to that of the original pool (mean $\rho = 0.712$–0.788). Because only 13 findings were compared, this decrease is descriptive. Third, higher recall did not correspond to better overall performance, as precision decreased at every K for every model, and F1 varied by model at $K = 1$ and $K = 5$ and was lower for all models at $K = 10$. These results indicate that part of the difficulty observed under the full retrieval pool reflects the composition of that pool, although not all of it.

The most plausible explanation for the recall gains is competition among candidate sentences. In the original pool, No Finding sentences accounted for 39.0% of the pool, whereas Pleural Other accounted for 0.4%. A sentence describing a rare finding therefore had to outrank a large number of more frequent sentences to appear among the top K. Balancing reduced the share of No Finding to 14.0% and increased that of Pleural Other to 7.3%, which provided more positive candidates for rare findings. The decrease in No Finding recall (0.119 to 0.187) is consistent with this explanation, as normal sentences were removed from competition with abnormal ones. Pool size alone does not account for the effect: at $K = 5$, random pools of equal size changed micro-recall only slightly and inconsistently (+0.013 for our model, +0.025 for CXR-ReDonE and -0.025 for CXR-RePaiR), whereas balanced-pool recall exceeded the random-pool mean by 0.07 to 0.08.

The pattern of finding-level gains is consistent with this account. Several of the largest increases occurred where baseline recall was low, for example Pneumothorax in our model (0.173 to 0.572) and Consolidation

Table 3: Change in finding-level recall at $K = 5$ under 30 balanced pools (mean $\pm$ SD of Δ = balanced $-$ original), compared with the 2.5th to 97.5th percentile range of the change under 50 random pools of the same size (in brackets).

Finding	Our model	CXR-RePaiR	CXR-ReDonE
No Finding	-0.187 ± 0.033 [-0.014, $+0.020$]	-0.119 ± 0.044 [$+0.017$, $+0.138$]	-0.130 ± 0.023 [$+0.016$, $+0.056$]
Enlarged Cardiomediastinum	$+0.337 \pm 0.034$ [-0.032, $+0.112$]	$+0.140 \pm 0.048$ [-0.142, $+0.019$]	$+0.208 \pm 0.027$ [-0.036, $+0.050$]
Cardiomegaly	$+0.162 \pm 0.055$ [-0.070, $+0.070$]	$+0.033 \pm 0.034$ [-0.178, -0.013]	$+0.159 \pm 0.037$ [-0.055, $+0.073$]
Lung Opacity	$+0.043 \pm 0.031$ [-0.050, $+0.052$]	$+0.040 \pm 0.044$ [-0.097, $+0.040$]	$+0.045 \pm 0.023$ [-0.026, $+0.062$]
Lung Lesion	$+0.178 \pm 0.031$ [-0.035, $+0.037$]	$+0.163 \pm 0.027$ [-0.034, $+0.058$]	$+0.165 \pm 0.020$ [-0.051, $+0.023$]
Edema	$+0.114 \pm 0.044$ [-0.055, $+0.074$]	$+0.098 \pm 0.026$ [-0.036, $+0.070$]	$+0.138 \pm 0.030$ [-0.077, $+0.072$]
Consolidation	$+0.185 \pm 0.040$ [-0.031, $+0.050$]	$+0.216 \pm 0.034$ [-0.115, $+0.062$]	$+0.203 \pm 0.034$ [-0.088, $+0.019$]
Pneumonia	$+0.143 \pm 0.025$ [-0.011, $+0.069$]	$+0.215 \pm 0.029$ [-0.014, $+0.114$]	$+0.176 \pm 0.029$ [-0.060, $+0.027$]
Atelectasis	$+0.154 \pm 0.044$ [-0.030, $+0.077$]	$+0.099 \pm 0.033$ [-0.071, $+0.072$]	$+0.111 \pm 0.024$ [-0.016, $+0.054$]
Pneumothorax	$+0.398 \pm 0.033$ [-0.023, $+0.157$]	$+0.166 \pm 0.015$ [-0.153, $+0.078$]	$+0.206 \pm 0.027$ [-0.020, $+0.055$]
Pleural Effusion	$+0.122 \pm 0.044$ [-0.047, $+0.071$]	$+0.040 \pm 0.046$ [-0.119, -0.006]	$+0.147 \pm 0.027$ [$+0.012$, $+0.086$]
Pleural Other	$+0.275 \pm 0.025$ [-0.053, $+0.053$]	$+0.209 \pm 0.032$ [-0.105, $+0.084$]	$+0.211 \pm 0.016$ [-0.079, 0.000]
Fracture	$+0.322 \pm 0.029$ [-0.053, $+0.091$]	$+0.160 \pm 0.057$ [-0.087, $+0.077$]	$+0.163 \pm 0.023$ [-0.026, $+0.045$]
Support Devices	$+0.015 \pm 0.031$ [-0.021, $+0.065$]	$+0.010 \pm 0.025$ [-0.038, $+0.054$]	$+0.047 \pm 0.026$ [$+0.008$, $+0.096$]

in CXR-ReDonE (0.180 to 0.383). The smallest changes occurred for findings with relatively high baseline recall, such as Support Devices (+0.010 to +0.047) and Lung Opacity (+0.040 to +0.045), and these were the combinations in which the balanced-pool change did not exceed the random-pool range in most pools. Because a finding was considered predicted when at least one of the top K sentences was positive, a pool containing more positive sentences for a finding increases its recall by construction. Recall under a balanced pool should therefore be interpreted as a property of a model-pool combination and not as direct evidence of improved image understanding.

The analysis of hard findings indicates that difficulty has both pool-dependent and persistent components. Under the original pool, Lung Lesion, Pleural Other and Fracture were among the five lowest-recall findings for all three models, whereas under the balanced pools only Lung Lesion remained hard for all three in every pool (mean recall 0.250 to 0.272). Fracture left the lowest five for our model in all 30 pools and Pleural Other in 12 of 30, with recall increases of +0.322 and +0.275, respectively. Both remained among the lowest five for CXR-RePaiR and CXR-ReDonE in most pools although their recall also increased (by 0.209 to 0.211 for Pleural Other and 0.160 to 0.163 for Fracture); because hard findings were defined by rank, remaining in the lowest five despite an absolute gain reflects gains in other findings as well. Lung Lesion remained hard

Table 4: Cross-model agreement in finding-level recall at $K = 5$ (Spearman ρ) under the original, balanced, and random retrieval pools. Original: 95% CI from 3,000 image-level bootstrap resamples. Balanced: mean ρ across 30 balanced pools, with the 2.5th to 97.5th percentile across pools in brackets. Random: mean $\pm$ SD of ρ across 50 random pools.

Findings	Model pair	ρ Original [95% CI]	ρ Balanced (mean [2.5th–97.5th pct.])	ρ Random (mean $\pm$ SD)
13 findings	Our model vs CXR-RePaiR	0.808 [0.736, 0.868]	0.528 [0.400, 0.687]	0.788 ± 0.069
(excluding	Our model vs CXR-ReDonE	0.753 [0.687, 0.835]	0.153 [−0.024, 0.392]	0.712 ± 0.093
No Finding)	CXR-RePaiR vs CXR-ReDonE	0.687 [0.599, 0.764]	0.577 [0.391, 0.749]	0.775 ± 0.078

despite a six-fold increase in its pool share (1.3% to 7.9%) and recall gains of 0.163 to 0.178. This suggests a limitation that pool share alone does not explain, which may be model-level or due to heterogeneity in how this finding is described in sentences, although the present design cannot distinguish between these possibilities.

The agreement results support the same interpretation. The original and random pools preserve the same label proportions in expectation and produced similar cross-model agreement ($\rho = 0.687$–0.808 and mean $\rho = 0.712$–0.788), which suggests that shared pool prevalence may account for much of the apparent consensus on which findings are difficult. When the proportions were equalized, model-specific behavior became more apparent. For example, our model showed larger gains than the other two models for Pneumothorax (+0.398 compared with +0.166 and +0.206) and Fracture (+0.322 compared with +0.160 and +0.163). One possible explanation, which was not tested here, is that region-word alignment benefits localized findings. The correlation between our model and CXR-ReDonE under the balanced pools was low (mean $\rho = 0.153$, range -0.024 to 0.392 across pools), but the decrease from the original pool was not statistically established when findings were resampled ($\Delta\rho = +0.597$, 95% CI -0.011 to $+1.280$; for our model with CXR-RePaiR, $\Delta\rho = +0.213$, 95% CI -0.194 to $+0.779$), so the lower agreement is better described as a descriptive observation than as evidence of disagreement. Correlations based on 13 findings are also sensitive to changes in individual ranks.

These results are consistent with evidence that retrieval metrics depend on the alternatives available to the model. Chen et al. [15] reported that recall decreased when cross-modal retrieval pools were enlarged with visually similar images, and the present results show the converse, in that reducing the number of competing normal sentences increased recall. Studies of information retrieval test collections have similarly shown that the way candidates are pooled can bias the assessment of retrieval systems [26]. In medical imaging, test-set composition [16] and the reference used to define pathology labels [17] have been shown to change measured performance, and in the latter case model rankings, without any change to the models. The present study identifies an analogous dependence on the retrieval pool in sentence-level CXR retrieval. The variation across findings under the original pool agrees with earlier reports: CXR-RePaiR was reported to reach an F1 of approximately 0.6 for pleural effusion and 0.15 for consolidation [5], and in the present evaluation its recall under the original pool was 0.650 and 0.256 for the same findings, although the metrics differ. Comparable spreads have been described for KAD [12] and ProtoCLIP [13]. The agreement between models under the original pool resembles the shared worst-performing classes reported by Gong et al. [18] for visual-semantic embedding networks. The present results add that this agreement weakened when pool proportions were changed by balancing but not when the pool was reduced by random sampling, which suggests that it partly reflects a shared pool distribution. Work on class imbalance and long-tailed chest radiograph classification [6, 23] has cautioned that aggregate metrics conceal uneven performance across findings, and the long-tailed

distribution of the sentence pool creates an analogous problem at inference time. Imbalance has also been studied in report generation itself [7], whereas the present study altered only the candidate pool.

Precision decreased for all models at all K (for example, from 0.3594 to 0.3051 for our model at $K = 5$). A plausible reason is that balancing shifts retrieval toward sentences describing less frequent findings, which are more often absent in a given image. This is consistent with the lower No Finding recall and with work showing that the proportion of positive cases affects precision and recall [27, 28]. At $K = 5$, F1 was stable for our model (0.3758 to 0.3755), higher for CXR-ReDonE (0.3815 to 0.3909) and lower for CXR-RePaiR (0.4623 to 0.4355), and at $K = 10$ it was lower for all three models. Micro-averaged scores are dominated by frequent findings, so substantial gains for rare findings have limited influence on them, which illustrates the value of finding-level analysis. Because the pool was modified at inference time and not the training data or the test-set class distribution, the results show that balancing changes the operating conditions of retrieval, not that it is equivalent to training-time resampling.

Three implications follow. First, studies of sentence-level retrieval should report the size and composition of the retrieval pool and compare models under identical pools, with size-matched random pools serving as a control. Methods that organize retrieval by topic [24] and reliability estimates grounded in a reference population [25] likewise depend on the distribution of the examples from which they retrieve, which supports reporting it. Second, comparing balanced and original pools can be used diagnostically: findings that improve with balancing, such as Fracture in our model, may be limited by the pool, whereas findings that remain hard under both pools, such as Lung Lesion, warrant closer examination as possible targets for model development. Third, pool curation can improve recall for rare findings but at the cost of precision and No Finding recall. In report generation this could lead to abnormal sentences being retrieved for normal studies, and systems that use balanced pools may therefore require safeguards such as similarity thresholds or abstention.

Limitations

Several limitations should be considered when interpreting these findings. The study shows that retrieval pool composition changes recall, apparent finding-level difficulty and, descriptively, cross-model agreement. It does not establish which property of the pool (class balance, proportion of normal sentences or co-occurrence of findings) is responsible, whether the lower agreement under balanced pools is statistically reliable, or that balanced pools improve report quality. First, the retrieval pool was built from the reports used to train our model, which may advantage that model, and comparisons across models should be interpreted accordingly. In addition, our model was trained by the authors, whereas the other two models were evaluated as released checkpoints with different training data, objectives and preprocessing (CXR-ReDonE, for example, was trained on impression sections, whereas the pool consists of findings sentences), so differences between models cannot be attributed to architecture alone. Second, only one balancing design (a cap of 450 sentences per finding) was evaluated. It altered class balance, the proportion of normal sentences and the co-occurrence of findings simultaneously, so their individual contributions cannot be separated, and intermediate degrees of balance were not examined. Balanced-pool results were averaged over 30 random draws of this single design. Third, labels for both sentences and images were obtained automatically, uncertain labels were excluded, and unmentioned findings were treated as negative. This is likely to affect absolute precision more than relative differences between pools, which share the same labeling. Fourth, correlations were computed over 13 findings, the confidence intervals for the original pool were obtained by resampling images and not findings, the definition of a hard finding (the five lowest-recall findings) was a fixed cutoff, and no correction was applied for the 42 comparisons against the random-pool range. Finally, evaluation used a single dataset

(CheXpert) with a pool derived from a different one (MIMIC-CXR), and measured retrieval against labels and not the quality of generated reports.

Future work

Future studies should vary pool size and composition independently and use graded levels of balance. They should relate finding-level recall directly to pool prevalence to quantify how much of the cross-model agreement it explains, and examine why findings such as Pleural Other and Fracture respond differently to balancing across models. Additional datasets, models and clinical findings should be included to establish whether the observed patterns generalize, and the effect of pool composition on the completeness and factual accuracy of generated reports should be assessed.

Conclusion

Retrieval pool composition influences finding-level recall and cross-model agreement in sentence-level CXR retrieval. Balanced pools increased recall for most findings in all three models, to a greater extent than size-matched random pools, but they reduced precision and did not consistently improve F1. Some retrieval difficulty is dependent on the pool, whereas some, such as that of Lung Lesion, persists across pools. Evaluations of retrieval models should therefore report pool composition and finding-level performance in addition to aggregate scores.

Ethical standards

This study did not involve any new data collection from human participants. All chest radiographs and associated clinical labels were obtained from publicly available, de-identified datasets (MIMIC-CXR and CheXpert), which received institutional approval and patient consent waivers at the time of their original release. No additional ethical approval was required for this secondary analysis.

Conflict of interest

The author declares no conflict of interest.

Data availability

The datasets used in this study are publicly available. Training data (MIMIC-CXR) were obtained from kaggle.com/datasets/zainabhalhoul/mimic-dataset, licensed under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Evaluation data (CheXpert) were obtained from kaggle.com/datasets/ashery/chexpert, licensed under CC0: Public Domain. Code and additional materials used in this study are available from the corresponding author upon reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ouis MY, Akhloufi MA. Deep learning for report generation on chest X-ray images. *Computerized Medical Imaging and Graphics*. 2024 Jan 1;111:102320.

[2] Li Y, Kong C, Zhao G, Zhao Z. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. *Artificial Intelligence Review*. 2025 Aug 21;58(11):344.

[3] Yang Y, You X, Zhang K, Fu Z, Wang X, Ding J, Sun J, Yu Z, Huang Q, Han W, Yu J. Spatio-temporal and retrieval-augmented modeling for chest x-ray report generation. *IEEE Transactions on Medical Imaging*. 2025 Mar 25;44(7):2892–905.

[4] Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, Fonseca EK, Lee HM, Abad ZS, Ng AY, Langlotz CP. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*. 2023 Sep 8;4(9).

[5] Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health* 2021 Nov 28 (pp. 209–219). PMLR.

[6] Mosquera C, Ferrer L, Milone DH, Luna D, Ferrante E. Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *European Radiology*. 2024 Dec;34(12):7895–903.

[7] Li W, Han G, Wu Y, Huang IC, Huang X. Joint Imbalance Adaptation for Radiology Report Generation. *Journal of Healthcare Informatics Research*. 2025 Dec;9(4):720–42.

[8] Jeong J, Tian K, Li A, Hartung S, Adithan S, Behzadi F, Calle J, Osayande D, Pohlen M, Rajpurkar P. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning* 2024 Jan 23 (pp. 978–990). PMLR.

[9] Zhang B, Jia C, Liu S, Schunkert H, Navab N. Semantic-aware chest x-ray report generation with domain-specific lexicon and diversity-controlled retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2025 Sep 19 (pp. 607–616). Cham: Springer Nature Switzerland.

[10] Ranjit M, Ganapathy G, Manuel R, Ganu T. Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine Learning for Healthcare Conference* 2023 Dec 22 (pp. 650–666). PMLR.

[11] You K, Gu J, Ham J, Park B, Kim J, Hong EK, Baek W, Roh B. CXR-CLIP: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2023 Oct 1 (pp. 101–111). Cham: Springer Nature Switzerland.

[12] Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*. 2023 Jul 28;14(1):4542.

[13] Kittler F, Bhat S, Maier A. ProtoCLIP: Prototype-Aligned Latent Refinement for Robust Zero-Shot Chest X-Ray Classification. *arXiv preprint* arXiv:2604.18444. 2026 Apr 20.

[14] Zhou Y, Faith TL, Xu Y, Leng S, Xu X, Liu Y, Goh RS. BenchX: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *Advances in Neural Information Processing Systems*. 2024 Dec 16;37:6625–47.

[15] Chen W, Yao L, Jin Q. Rethinking benchmarks for cross-modal image-text retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2023 Jul 19 (pp. 1241–1251).

[16] Till T, Scherkl M, Stranger N, Singer G, Hankel S, Flucher C, Hržić F, Štajduhar I, Tschauner S. Impact of test set composition on AI performance in pediatric wrist fracture detection in X-rays. *European Radiology*. 2025 Nov;35(11):6853–64.

[17] Fytas P, Selby I, Karner C, Babar J, Baker S, Beckford J, Sadler TJ, Shahipasand S, Thavakumar A, Chen JL, Sawer A. Rethinking Clinical Relevance in Chest X-ray Machine Learning: How Evaluation References Define Performance. *arXiv preprint* arXiv:2607.26333. 2026 Jul 28.

[18] Gong Y, Cosma G, Fang H. On the limitations of visual-semantic embedding networks for image-to-text information retrieval. *Journal of Imaging*. 2021 Jul 26;7(8):125.

[19] MacAvaney S, Feldman S, Goharian N, Downey D, Cohan A. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*. 2022 Mar 18;10:224–39.

[20] Hendriksen M, Zhang S, Reinanda R, Yahya M, Meij E, de Rijke M. Benchmark granularity and model robustness for image-text retrieval: a reproducibility study. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2025 Jul 13 (pp. 3183–3193).

[21] Ramesh V, Chi NA, Rajpurkar P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health* 2022 Nov 22 (pp. 456–473). PMLR.

[22] Ji Z, Shaikh MA, Moukheiber D, Srihari SN, Peng Y, Gao M. Improving joint learning of chest x-ray and radiology report by word region alignment. In *International Workshop on Machine Learning in Medical Imaging* 2021 Sep 21 (pp. 110–119). Cham: Springer International Publishing.

[23] Holste G, Zhou Y, Wang S, et al. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Medical Image Analysis*. 2024;97:103224. doi:10.1016/j.media.2024.103224.

[24] Zhao J, Zhou Y, Chen Z, Fu H, Wan L. Topicwise separable sentence retrieval for medical report generation. *IEEE Transactions on Medical Imaging*. 2025;44(3):1505–1517. doi:10.1109/TMI.2024.3507076.

[25] Alanazi MF, Shakhreet BZ, Asiri HAA, et al. CBEC: A simple retrieval-based framework for population-grounded prediction reliability estimation in chest X-ray diagnosis. *Scientific Reports*. 2026. doi:10.1038/s41598-026-66377-5.

[26] Buckley C, Dimmick D, Soboroff I, Voorhees EM. Bias and the limits of pooling for large collections. *Information Retrieval*. 2007;10(6):491–508. doi:10.1007/s10791-007-9032-x.

[27] Erenel Z, Altınçay H. Improving the precision-recall trade-off in undersampling-based binary text categorization using unanimity rule. *Neural Computing and Applications*. 2013;22:83–100. doi:10.1007/s00521-012-1056-5.

[28] Shang H, Langlois JM, Tsioutsiouliklis K, Kang C. Precision/Recall on Imbalanced Test Data. In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research. 2023;206:9879–9891.