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.

Zero-shot performance of a concept-grounded chest radiograph foundation model depends on the evaluation dataset: a preregistered external evaluation of CLEAR and CheXzero

Ismail Dahiru

ismaildahiru957@gmail.com

Mewar University

Buhari Aminu Adam

Mewar University

Surajo Nuhu Umar

Shobhit Institute of Engineering & Technology

Bachcha Lal Pal

Mewar University

Research Article

Keywords: foundation models, chest radiography, zero-shot classification, external validation, concept bottleneck, algorithmic fairness

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11444323/v1

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

Additional Declarations: No competing interests reported.

Abstract

CLEAR, a concept-grounded vision-language model presented as auditable, scores chest radiographs through 368,294 radiological concepts. It reportedly outperformed CheXzero in zero-shot classification, but only on developer-chosen test sets, without a subgroup analysis or a systematic check of the concepts behind its zero-shot predictions. In a preregistered comparison on the NIH ChestX-ray14 test split (25,592 images, 2,797 patients), macro-AUROC over 14 findings was 0.695 for CheXzero, 0.635 for CLEAR's concept route and 0.700 for direct scoring with CLEAR's encoder. Neither model showed evidence of a sex gap (female minus male: CheXzero + 0.005, 95% CI − 0.015 to + 0.022; concept route − 0.008, − 0.029 to + 0.011) or an age gap under any of four prompt templates. Prompt wording alone reversed the ranking of CLEAR's direct route against CheXzero (− 0.038 to + 0.031). The concept space was nearly collapsed (participation ratio 2.95, against 11.18 for the image features), and one held-out-selected concept outperformed the concept route for every finding. On 15,000 VinDr-CXR radiographs the concept route led CheXzero by 0.057 (0.050 to 0.064), in the reported direction, but by only 0.023 where sex was recorded and 0.009 where age was recorded, the images carrying most of the disease. On the Indiana University collection, one of CLEAR's own test sets, it trailed CheXzero by 0.015 (0.007 to 0.024), using report indexing terms as labels. CLEAR's zero-shot advantage depended on the dataset, and the concept step added no accuracy over the image encoder. Developer-reported zero-shot gains and concept explanations need independent testing before health systems adopt them.

1 Introduction

Vision-language models trained on paired chest radiographs and reports can classify findings without task-specific labels by comparing an image with text prompts [1–3]. CheXzero, trained on MIMIC-CXR [4], performed comparably to three board-certified radiologists on five CheXpert findings it had not been explicitly trained to classify [2]. CLEAR goes a step further [3]. It first expresses the image as similarities to 368,294 radiological observations mined from reports, projects that vector into the embedding space of a large language model, and only then compares it with the prompt. The authors present this concept route as both more accurate and auditable, since each prediction can be traced to named observations. They report that CLEAR outperformed CheXzero in zero-shot classification on VinDr-CXR (mean AUROC 78.2% against 75.0%) and PadChest, and on the Indiana University collection was better on 10 findings and worse on 2 [3].

We found no published independent evaluation of these results. They come from test sets the developers selected and processed, and medical imaging models can lose accuracy when the data source changes [5]; in one review only about a quarter of machine learning papers in health used more than one dataset [6]. Two further questions are rarely asked of zero-shot models. One is prompt sensitivity: prompt choice alone moved zero-shot AUROC on the CheXpert validation set by more than 25 points in one study [7] and by about 3 points on ChestX-ray14 in another [8], yet published comparisons usually report a single template. The other is whether accuracy is equal across patient groups. Supervised chest radiograph classifiers underdiagnosed female and younger patients on MIMIC-CXR and

CheXpert [9], and protected attributes are encoded in the features of a chest radiograph foundation model [10] and of supervised disease classifiers [11]. CheXzero underdiagnosed female and younger patients on several public datasets at a fixed threshold [12], and six CLIP-based models showed gaps by age on MIMIC-CXR, with more even results by sex and race [13]. CLEAR's report did not include a subgroup analysis [3].

Auditability raises a separate question. Radiologists in the CLEAR study rated the leading concepts of supervised linear probes [3], and zero-shot predictions were illustrated only with example concept plots. If the concepts behind such predictions have little to do with the finding, the explanation misleads its reader. Because zero-shot models need no local labels, a health system could adopt one on the strength of the developers' benchmarks alone. We therefore carried out a preregistered external evaluation of CLEAR and CheXzero. The registered primary analysis tested whether zero-shot performance on NIH ChestX-ray14 [14], a dataset used neither to train nor to evaluate CLEAR, differed by sex and age. Two amendments, filed on OSF before the corresponding runs, added a test of prompt robustness, an audit of the concept route, and a replication on VinDr-CXR [15]. We also repeated the comparison on the Indiana University collection [16], one of CLEAR's own external test sets.

2 Methods

2.1 Registration and reporting

This was a retrospective study of public datasets. The protocol was registered on the Open Science Framework on 1 October 2026 (https://osf.io/dsbem; OSF Preregistration template), before any performance metric was computed on the test images; by then we had inspected the labels and subgroup counts and run a 100-image technical check of both models. The registration fixed the cohort, eligibility rule, hypotheses, primary metric and bootstrap procedure (Fig. 1). Amendment 1 (prompt robustness and concept audit) and Amendment 2 (VinDr-CXR) were filed in the registration's OSF project (https://osf.io/ty7d5) on 3 October 2026, after the primary results were known but before either analysis was run. Four registered analyses could not be done as planned (reasons in Online Resource 1). Reporting follows the CLAIM 2024 checklist [17] where applicable.

2.2 Data

NIH ChestX-ray14, the extended version of ChestX-ray8 [14], contains 112,120 frontal radiographs from 30,805 patients, with 14 findings mined from the reports by natural language processing [5, 9]. We used the official test split, which is separated from the training split by patient, after excluding four images with a recorded age above 100 years: 25,592 images from 2,797 patients (1,240 female, 1,557 male). Its size was set by the split, and no power calculation was made. A separate 10% of the training patients (2,800 patients, 8,441 images; random seed 2026) was held out to set decision thresholds and to select concepts in the audit, and was used for nothing else.

The Indiana University collection [16] was taken from its public Kaggle copy: 7,466 images (3,818 frontal) from 3,851 reports. Labels came from the report's Problems field, the indexing terms assigned to each report.

VinDr-CXR [15] was used through the training split of the Kaggle VinBigData competition, under its rules for non-commercial academic research. The CLEAR report used the official 3,000-image test split with 21 findings [3]; its labels are not in the Kaggle release, so ours is a separate test, not a repeat of that benchmark. Each of the 15,000 DICOM images was read by three radiologists. A finding was positive when at least two marked it, and positivity by any one radiologist served as a sensitivity analysis. We kept the 13 findings with a natural-language name (Online Resource 1). Sex and age came from the DICOM headers, with age treated as missing when absent, zero or above 100 years.

2.3 Models

CheXzero is a CLIP ViT-B/32 model [1] trained further on MIMIC-CXR image-report pairs [2]. As in the original report, we used the mean prediction of its 10 released checkpoints (file names in Online Resource 2), with images preprocessed as in the released code (Online Resource 1).

CLEAR uses a DINOv2 ViT-B/14 image encoder at 448 pixels [18], trained on 873,342 image-report pairs from MIMIC-CXR, CheXpert-Plus and ReXGradient [3]; code and weight versions are in Online Resource 2. Non-square images were padded to 448 pixels and normalised with CLEAR's constants. Because the CLEAR paper resizes images to 448 x 448 without padding [3], the VinDr-CXR scoring was repeated that way (post hoc). Computing details are in Online Resource 1.

2.4 Zero-shot scoring

Each finding was scored with a positive prompt, the lower-case finding name (for example "pleural effusion"), and a negative prompt formed by prefixing "no ", the same for both models. The score was the softmax of the two cosine similarities, as in CheXzero [2] and CLEAR [3].

As in its code release (Fig. 1a), CLEAR's concept route compares the L2-normalised image embedding with the CLIP embeddings of all 368,294 concepts, multiplies the similarity vector by the concepts' SFR-Embedding-Mistral embeddings [19], normalises the result and compares it with the SFR embeddings of the two prompts. Our implementation matched the authors' pipeline on real test images (Online Resource 1).

During the first scoring run we found that the released SFR concept embeddings could be reproduced only with an instruction prefix (mean cosine 0.999998, against 0.899 without it; wording in Online Resource 1). The CLEAR paper says the concepts were embedded with a task instruction but not whether the class prompts were [3], and the released zero-shot script embeds prompts without it. The script's plain prompts are therefore our primary method, and prompts with the prefix a secondary variant, added after registration and disclosed in Amendment 1. CLEAR's direct route, a further secondary score, compares the same image embedding with CLEAR's own text-encoder embeddings of the two prompts.

The two routes differ in both the concept step and the prompt encoder, so their difference measures the two together.

2.5 Registered outcomes and statistical analysis

The primary metric was AUROC, computed exactly from the Mann-Whitney statistic [20] and averaged over findings (macro-AUROC). Subgroups were sex and four age bands (0 to 18, 18 to 40, 40 to 60 and over 60 years), with 40 to 60 as the reference. A finding entered a comparison only if every subgroup in it had at least 30 positive images from at least 10 patients. Applied to labels alone, this rule left 14 findings for sex and 11 for age (emphysema, fibrosis and hernia were excluded).

The registered hypotheses were that CheXzero (H1) and CLEAR (H2) show a sex or age gap in macro-AUROC, and that the two models differ in the size of these gaps (H3). Uncertainty came from 1,000 bootstrap resamples of patients (seed 2026), paired across subgroups and models [21]. Intervals are 95% for sex and 98.33% for the three age contrasts (Bonferroni). Secondary analyses repeated the sex comparison on posteroanterior images (where the age comparison had two eligible findings and was descriptive only) and compared false-negative rates at the Youden threshold set on the held-out patients.

2.6 Amendment 1: prompt robustness and concept audit

Prompt robustness was tested with four templates applied to every finding and model: the registered T1, "{f}" / "no {f}", and T2 to T4 (wording in Fig. 1b). A subgroup conclusion was called prompt-robust if every contrast had the same verdict (interval excluding zero or not) under all four templates.

For the concept audit, the score difference between the positive and negative prompt was written as a sum of per-concept contributions. A concept counted as relevant if its text matched a pattern for the finding listed in Amendment 1 (for cardiomegaly these included "cardiomegal", "heart size", "enlarged heart" and "cardiac silhouette"). Explanation relevance was the share of relevant concepts among the 10 largest contributors per image, compared across subgroups among true positives. For each finding we also chose the relevant concept with the highest AUROC on the held-out patients and evaluated it on the test split; this probe uses labels, so it shows where information sits and is not a zero-shot method. Finally, we measured the spread of 2,000 random test images in CLEAR's image space, its concept space and CheXzero's image space (checkpoint 1) by the mean pairwise cosine similarity and the participation ratio, the effective number of dimensions in use [22].

2.7 External datasets

The Indiana comparison was fixed in the analysis notebook before it was run and disclosed in Amendment 1. Its primary estimate was CLEAR concept minus CheXzero macro-AUROC over every Problems term with at least 20 positive reports, with a 95% interval from 1,000 report-level bootstrap resamples. The result was to be called consistent with the CLEAR report if the difference was positive and CLEAR won more findings than it lost, inconsistent if it was negative with an interval excluding zero, and inconclusive otherwise.

Amendment 2 applied the same rule to VinDr-CXR over the 13 findings with at least 30 positive images, resampling images because the release has no patient identifier. The registered sex and age contrasts were repeated as exploratory analyses on images with the relevant header field. Because fewer than 30 positive images per finding were available for patients aged 18 or younger, the age comparison used 18 to 40 and over 60 against 40 to 60, as the amendment specified.

2.8 Post hoc analyses

Post hoc analyses, including one disclosed in Amendment 1, are listed in Online Resource 1 and labelled as such where reported.

2.9 Use of AI and writing tools

We used ChatGPT (OpenAI) and Claude (Anthropic) through their apps in September and October 2026, with default settings, to help write and debug code, check numbers against run logs, and edit text; Grammarly for grammar and paraphrasing; and Zotero for references. Versions are in Online Resource 1. The authors reviewed all output and take full responsibility for the article.

3 Results

3.1 Cohorts

Table 1 describes the cohorts, and Online Resource 3 gives disease prevalence and the share of anteroposterior images in each NIH subgroup. The NIH test split had about nine images per patient. In VinDr-CXR, sex was missing from 51% of headers and age from 75%.

Table 1
Cohorts. Images (patients or reports) per group; VinDr-CXR demographics from DICOM headers.

Group	NIH test	NIH held-out	Indiana	VinDr-CXR
All	25,592 (2,797 patients)	8,441 (2,800 patients)	7,466 (3,851 reports)	15,000
Female	10,713 (1,240)	3,750 (1,302)		3,514
Male	14,879 (1,557)	4,691 (1,498)		3,840
Sex missing	0	0		7,646
Age 0 to 18	1,153 (147)	510 (172)		113
Age 18 to 40	7,719 (967)	2,166 (860)		780
Age 40 to 60	11,151 (1,179)	3,784 (1,265)		1,337
Age over 60	5,569 (624)	1,981 (617)		1,530
Age missing	0	0		11,240
Frontal / anteroposterior share	all frontal; 56.6% AP	all frontal; 34.9% AP	3,818 frontal	
Label source	text-mined, 14 findings	as test	Problems field	3 radiologists, 13 findings

3.2 Zero-shot accuracy on NIH ChestX-ray14

Macro-AUROC over the 14 findings was 0.695 (95% CI 0.685 to 0.704) for CheXzero, 0.635 (0.624 to 0.644) for CLEAR's concept route, 0.700 (0.690 to 0.709) for CLEAR's direct route and 0.640 (0.630 to 0.650) for the concept route with instructed prompts (Fig. 2; Online Resource 4). The concept route trailed CheXzero by 0.060 (paired 95% CI 0.051 to 0.069) and the direct route by 0.065 (0.059 to 0.072), while the direct route did not differ from CheXzero (+ 0.006, − 0.002 to + 0.013). CheXzero's result lay between the mean AUROC reported for it on PadChest (0.668) and VinDr-CXR (0.750) in the CLEAR report [3].

The losses were concentrated in a few findings. Cardiomegaly reached 0.614 with the concept route, against 0.821 with the direct route from the same image embedding and 0.835 with CheXzero, and hernia (0.377) and emphysema (0.495) were at or below chance. The concept route matched or exceeded the direct route for atelectasis, infiltration, nodule and fibrosis.

The similarity of each finding's two prompts did not explain these failures (post hoc; Online Resource 1).

3.3 Sex and age subgroups

None of the registered hypotheses was supported (Table 2, Fig. 3). The female minus male difference in macro-AUROC was + 0.005 (95% CI − 0.015 to + 0.022) for CheXzero and − 0.008 (− 0.029 to + 0.011) for

CLEAR's concept route. All six age contrasts against 40 to 60 years had 98.33% intervals including zero, with estimates from − 0.013 to + 0.023. The difference between the two models' gaps (H3) was − 0.013 (− 0.029 to + 0.006) for sex and within 0.02 of zero for every age contrast, with all intervals including zero. Sex gaps larger than about 0.03 AUROC are therefore unlikely for either model on this dataset.

Table 2
Registered subgroup analysis on NIH (macro-AUROC; differences with 95% CI for sex and 98.33% CI for age; 14 findings for sex, 11 for age).

Contrast	CheXzero	CLEAR concept	Difference between models (H3)
Female, male	0.697, 0.692	0.630, 0.638	
Female minus male	+ 0.005 (− 0.015 to + 0.022)	−0.008 (− 0.029 to + 0.011)	−0.013 (− 0.029 to + 0.006)
0 to 18 minus 40 to 60	−0.013 (− 0.054 to + 0.026)	+ 0.006 (− 0.042 to + 0.057)	+ 0.019 (− 0.023 to + 0.064)
18 to 40 minus 40 to 60	+ 0.013 (− 0.010 to + 0.034)	+ 0.023 (− 0.001 to + 0.048)	+ 0.010 (− 0.007 to + 0.026)
Over 60 minus 40 to 60	−0.002 (− 0.023 to + 0.017)	−0.008 (− 0.033 to + 0.016)	−0.007 (− 0.025 to + 0.012)
Posteroanterior only, female minus male	−0.005 (− 0.026 to + 0.017)	−0.016 (− 0.039 to + 0.008)	
False-negative rate, female minus male	−0.010 (− 0.038 to + 0.020)	−0.008 (− 0.038 to + 0.020)	

Results were the same on posteroanterior images (Table 2), and false-negative rates at the held-out threshold did not differ by sex or age for either model. Among the secondary scores, only CLEAR's direct route showed an interval excluding zero, for 18 to 40 against 40 to 60 years (+ 0.023, 98.33% CI + 0.003 to + 0.043).

3.4 Prompt wording

Overall macro-AUROC moved with the template by 0.020 for CheXzero (0.674 to 0.695), 0.046 for the concept route (0.622 to 0.667) and 0.080 for the direct route (0.636 to 0.717), and every shift from T1 had a paired interval excluding zero (Fig. 4a; Online Resource 5). The concept route stayed behind CheXzero under all four templates, by 0.020 to 0.064. The direct route was level with CheXzero under T1 and T2, behind it under T3 (− 0.038, 95% CI − 0.046 to − 0.031) and ahead of it under T4 (+ 0.031, + 0.023 to + 0.038), so their ranking depended on the prompt wording (Fig. 4b).

The subgroup conclusions for CheXzero and the concept route were prompt-robust: no sex or age contrast had an interval excluding zero under any template. The direct route's single positive age contrast (18 to 40 against 40 to 60 years) appeared under T1 only.

3.5 Concept audit

A single relevant concept, chosen on held-out patients, outperformed the concept route on the test split for all 14 findings and the direct route for 12 (Fig. 5a), for example cardiomegaly (0.886 against 0.614 for the concept route), emphysema (0.897 against 0.495) and hernia (0.854, 95% CI 0.776 to 0.919, against 0.377). For emphysema and pneumothorax the chosen sentence described subcutaneous emphysema, a neighbouring entity.

Explanation relevance among true positives ranged from 0.79 for atelectasis to 0.03 for mass and was below 0.15 for infiltration, pneumothorax, emphysema and mass (Fig. 5b; Online Resource 6). For emphysema and hernia, the concepts contributing most to true-positive predictions were normal-lung statements such as "lungs are clear without focal consolidation, effusion, or pneumothorax". For cardiomegaly the leading concepts were relevant ("enlargement of the heart"), but false negatives also drew on "enlargement of the thyroid" and "enlargement of the thoracic aorta". Relevance did not differ by sex (− 0.012, 95% CI − 0.046 to + 0.016) or age (all intervals including zero).

The concept representation was nearly collapsed. Random test images had a mean pairwise cosine similarity of 0.98 and a participation ratio of 2.95 in the concept space, against 0.30 and 11.18 in CLEAR's image space and 0.70 and 10.29 in CheXzero's.

In the three randomly drawn true positives of Fig. 6 (post hoc), the ten leading concepts were all normal-lung statements for emphysema but mostly named the finding for cardiomegaly and hernia, although hernia true positives were far less relevant on average (Online Resource 6). In each image they made up a small part of the total contribution, which for the hernia image pointed the opposite way.

3.6 VinDr-CXR

On VinDr-CXR, macro-AUROC over 13 findings was 0.817 (95% CI 0.809 to 0.824) for CheXzero, 0.873 (0.868 to 0.879) for CLEAR's concept route and 0.878 (0.872 to 0.884) for its direct route (Fig. 7; Online Resource 7). The concept route led CheXzero by 0.057 (0.050 to 0.064) and won 9 of 13 findings, meeting the registered definition of consistency with the CLEAR report; with labels from any single radiologist the lead was 0.069 (0.063 to 0.075). The concept route trailed the direct route by 0.005 (0.001 to 0.009), and cardiomegaly was again lower with the concept route (0.869) than with the direct route (0.917) or CheXzero (0.910).

The images with recorded demographics differed from the rest (post hoc): 48% of images with sex recorded had none of the 13 findings, against 93% of those without. In the 7,354 images with sex recorded, the concept route led CheXzero by 0.023 (0.016 to 0.031) and won 6 of 13 findings. In the 3,760 images with age recorded the lead was 0.009 (0.000 to 0.019), and in the mostly normal images without recorded sex it was 0.098 (0.080 to 0.117; 7 eligible findings). The direct route led CheXzero in every subset, by 0.032 to 0.062. Scoring with CLEAR's own resizing instead of padding raised the concept route by 0.006 (95% CI 0.004 to 0.008) and left this pattern unchanged (Online Resource 1).

In the exploratory VinDr-CXR subgroup analysis (10 eligible findings for sex, 6 for age), all three scores were higher for patients aged 18 to 40 than for those aged 40 to 60, with 98.33% intervals excluding zero (CheXzero + 0.066, concept route + 0.069, direct route + 0.038; Online Resource 1); for CheXzero and the concept route the gap persisted when each finding was left out in turn. CheXzero also showed a small female advantage (+ 0.017, 95% CI + 0.002 to + 0.032). Prevalence rose steeply with age (aortic enlargement 0.9%, 26.5% and 68.6% in the three bands).

3.7 Indiana University collection

Over the 51 Problems terms with at least 20 positive reports, macro-AUROC was 0.701 (95% CI 0.692 to 0.709) for CheXzero, 0.686 (0.677 to 0.695) for CLEAR's concept route and 0.688 (0.679 to 0.697) for its direct route (Fig. 7). The concept route trailed CheXzero by 0.015 (0.007 to 0.024) and won 22 of 51 terms, which the fixed rule classes as inconsistent with the CLEAR report. Every CLEAR variant trailed CheXzero, on all images and on frontal images alone. The terms included anatomical ones such as "Lung" and "Spine"; on the 14 that match NIH or CheXpert findings, the concept route trailed by 0.046 and won 4 (post hoc; Online Resource 1).

4 Discussion

On NIH ChestX-ray14, CLEAR's concept route was less accurate than CheXzero and than CLEAR's own direct route. On VinDr-CXR it was more accurate than CheXzero, as originally reported, but by a wide margin only among mostly normal images, and on Indiana, with our labels, it trailed CheXzero. Under the registered template the direct route was level with CheXzero on NIH, ahead on VinDr-CXR and slightly behind on Indiana, and on no dataset did the concept route improve on it.

The audit suggests why. The concept vector of almost every image pointed in nearly the same direction (mean cosine 0.98), so classification rested on small deviations from a shared profile. In that regime, broad concepts that are close to every image, such as normal-lung statements, can dominate the explanation of an abnormal image, as they did on average for emphysema and hernia. The information needed was not missing: for every finding, one held-out-selected concept separated cases from controls better than the full route and better than CheXzero. This points to the step in which 368,294 similarities are aggregated through text embeddings into one vector. Anisotropic embedding spaces of this kind are known in language models [23], and on general images LLM-derived concept descriptors performed no better than random words [24]. In a concept bottleneck [25], concepts are meant to carry the decision; for several findings here, the concepts that drove the score had little to do with the finding. This does not contradict the 89.8% relevance rating in the CLEAR report [3], which concerned linear probes.

This study cannot separate the sources of the dataset dependence: label construction (text-mined labels in NIH, report indexing terms in Indiana, radiologist marks in VinDr-CXR), case mix, image format and acquisition. As the CLEAR authors noted in correspondence, NIH was not one of their benchmarks and our Indiana labels differ from theirs. The VinDr-CXR subsets show how far case mix can move a comparison: the same models and labels gave a lead of 0.098 in the mostly normal images and 0.023 in

the rest, although the two parts also differed in eligible findings and may differ in source hospital. A zero-shot advantage measured on one external set should not be assumed to transfer.

Prompt wording had a large effect. The direct route's AUROC changed by 0.080 between templates, and its ranking against CheXzero reversed. Prompt sensitivity has been reported before [7, 8], and the CLEAR authors found AUROC variation of 0.014 to 0.029 across six prompt pairs on their external sets [3]; on NIH the concept route varied by 0.046 across our four templates. Zero-shot comparisons should therefore report more than one template or fix the template before seeing results, as we did for T1.

On NIH neither model showed a sex or age gap under any template. Earlier audits led us to expect gaps for CheXzero [12, 13], but they measured something else. Yang et al. scored the "no finding" label on all 112,120 NIH images at a threshold set on validation data and found female and younger patients more often wrongly labelled as having no finding [12]. Seyyed-Kalantari et al., with the same label on NIH and a supervised classifier, found male patients and those over 80 years, a small group, most often underdiagnosed [9]. Sourget et al. examined subgroups on MIMIC-CXR only [13]. We measured macro-AUROC over the eligible findings on the official test split, which had only 25 patients over 80, and our false-negative analysis at a held-out threshold, the closest to those measures, also showed no gap. On VinDr-CXR the younger group had higher AUROC for every model, but with several-fold differences in prevalence between age bands and three quarters of ages missing, case mix is the likely explanation. Model bias cannot be ruled out: rebalancing disease prevalence in the test set narrowed some subgroup gaps of supervised classifiers but did not remove them [11].

The study has limitations. NIH labels were mined from text and contain errors [26], which lowers measured AUROC, unevenly across findings; in one radiologist review, most images labelled emphysema showed subcutaneous rather than pulmonary emphysema [26], which may explain why subcutaneous-emphysema sentences were selected for that label. We could not obtain the radiologist-adjudicated labels released for 1,962 NIH test images and four findings [27], so the registered check against them was not run. Indiana labels were built from indexing terms, not from CLEAR's unpublished label set. VinDr-CXR was evaluated on its training split with an image-level bootstrap, because patient identifiers are not released, and its subgroup analyses rest on incomplete headers. None of the three datasets records race, which earlier CheXzero audits examined [12, 13]. We evaluated only the released concept library and LLM embedding. The relevance patterns fixed in Amendment 1 miss some synonyms, such as "enlargement of the heart" for cardiomegaly, so the relevance shares are approximate. We studied zero-shot classification only; CLEAR's linear-probe results [3] were outside our scope.

5 Conclusions

In this preregistered external evaluation, CLEAR's zero-shot advantage over CheXzero depended on the dataset. Its concept route lost accuracy relative to the image encoder it is built on, its concept space was nearly collapsed, and its explanations for several findings were dominated by irrelevant concepts. Neither CheXzero nor CLEAR's concept route showed evidence of a sex or age gap on NIH ChestX-ray14

under any of four prompt templates. Before adopting a zero-shot chest radiograph model, health systems should require results on independent data under more than one prompt template, and evidence that its explanations relate to the finding predicted.

Declarations

Competing interests:

the authors declare no competing interests.

Ethics approval:

not required. All three datasets are public and de-identified.

Consent to participate:

not applicable.

Consent for publication:

not applicable.

Funding:

none.

Author Contribution

I.D. designed the study, wrote the analysis code, ran the analyses and wrote the first draft. B.A.A. contributed to the study design, checked the analyses and results, and critically reviewed the manuscript. S.N.U. contributed to the literature review and the interpretation of the results and reviewed and edited the manuscript. B.L.P. contributed to the conception and design of the study and the interpretation of the results, supervised the study, and reviewed and edited the manuscript. All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

Data Availability

NIH ChestX-ray14 was provided by the NIH Clinical Center and is available at https://nihcc.app.box.com/v/ChestXray-NIHCC. The Indiana University collection and VinDr-CXR are available from their providers. The registration is on the Open Science Framework (https://osf.io/dsbem) and its amendments are in the linked project (https://osf.io/ty7d5), and the per-finding results are released with the code. The analysis code, run logs and per-image scores are available on GitHub (https://github.com/ismaildahiru957/clear-chexzero-external-evaluation), archived on Zenodo (https://doi.org/10.5281/zenodo.23453344).

Code availability:

the analysis code, run logs and per-image scores are available on GitHub (https://github.com/ismaildahiru957/clear-chexzero-external-evaluation), archived on Zenodo (https://doi.org/10.5281/zenodo.23453344) and linked from the Open Science Framework project (https://osf.io/ty7d5).

References

1. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: Proc 38th Int Conf Mach Learn (ICML). PMLR 139:8748–8763
2. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P (2022) Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 6:1399–1406. https://doi.org/10.1038/s41551-022-00936-9
3. Han T, Wu R, Tian Y, et al (2026) CLEAR: an auditable foundation model for radiology grounded in clinical concepts. Nat Biomed Eng. https://doi.org/10.1038/s41551-026-01741-4
4. Johnson AEW, Pollard TJ, Berkowitz SJ, et al (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6:317. https://doi.org/10.1038/s41597-019-0322-0
5. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 15:e1002683. https://doi.org/10.1371/journal.pmed.1002683
6. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M (2021) Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 13:eabb1655. https://doi.org/10.1126/scitranslmed.abb1655
7. Pellegrini C, Keicher M, Özsoy E, Jiraskova P, Braren R, Navab N (2023) Xplainer: from X-ray observations to explainable zero-shot diagnosis. In: Medical Image Computing and Computer Assisted Intervention, MICCAI 2023. Lect Notes Comput Sci 14224:420–429. https://doi.org/10.1007/978-3-031-43904-9_41
8. Lai H, Yao Q, Jiang Z, et al (2024) CARZero: cross-attention alignment for radiology zero-shot classification. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 11137–11146.

https://doi.org/10.1109/CVPR52733.2024.01059
9. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27:2176–2182. https://doi.org/10.1038/s41591-021-01595-0
10. Glocker B, Jones C, Roschewitz M, Winzeck S (2023) Risk of bias in chest radiography deep learning foundation models. Radiol Artif Intell 5:e230060. https://doi.org/10.1148/ryai.230060
11. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M (2024) The limits of fair medical imaging AI in real-world generalization. Nat Med 30:2838–2848. https://doi.org/10.1038/s41591-024-03113-4
12. Yang Y, Liu Y, Liu X, et al (2025) Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv 11:eadq0305. https://doi.org/10.1126/sciadv.adq0305
13. Sourget T, Restrepo D, Hudelot C, Ferrante E, Christodoulidis S, Vakalopoulou M (2025) Fairness and robustness of CLIP-based models for chest X-rays. arXiv 2507.21291. https://doi.org/10.48550/arXiv.2507.21291
14. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), pp 3462–3471. https://doi.org/10.1109/CVPR.2017.369
15. Nguyen HQ, Lam K, Le LT, et al (2022) VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data 9:429. https://doi.org/10.1038/s41597-022-01498-w
16. Demner-Fushman D, Kohli MD, Rosenman MB, et al (2016) Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 23:304–310. https://doi.org/10.1093/jamia/ocv080
17. Tejani AS, Klontzas ME, Gatti AA, et al (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell 6:e240300. https://doi.org/10.1148/ryai.240300
18. Oquab M, Darcet T, Moutakanni T, et al (2024) DINOv2: learning robust visual features without supervision. Trans Mach Learn Res. https://openreview.net/forum?id=a68SUt6zFt
19. Meng R, Liu Y, Joty SR, Xiong C, Zhou Y, Yavuz S (2024) SFR-Embedding-Mistral: enhance text retrieval with transfer learning. Salesforce AI Research Blog. https://www.salesforce.com/blog/sfr-embedding/. Accessed 4 Oct 2026
20. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. https://doi.org/10.1148/radiology.143.1.7063747
21. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
22. Gao P, Trautmann E, Yu B, et al (2017) A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv 214262. https://doi.org/10.1101/214262
23. Ethayarajh K (2019) How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proc 2019 Conf Empir Methods Nat Lang

Process and 9th Int Jt Conf Nat Lang Process (EMNLP-IJCNLP), pp 55–65. https://doi.org/10.18653/v1/D19-1006

24. Roth K, Kim JM, Koepke AS, Vinyals O, Schmid C, Akata Z (2023) Waffling around for performance: visual classification with random words and broad concepts. In: Proc IEEE/CVF Int Conf Comput Vis (ICCV), pp 15700–15711. https://doi.org/10.1109/ICCV51070.2023.01443
25. Koh PW, Nguyen T, Tang YS, et al (2020) Concept bottleneck models. In: Proc 37th Int Conf Mach Learn (ICML). PMLR 119:5338–5348
26. Oakden-Rayner L (2020) Exploring large-scale public medical image datasets. Acad Radiol 27:106–112. https://doi.org/10.1016/j.acra.2019.10.006
27. Majkowska A, Mittal S, Steiner DF, et al (2020) Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology 294:421–431. https://doi.org/10.1148/radiol.2019191293