nature portfolio

Corresponding author(s): Tianyu Han

Last updated by author(s): May 1, 2026

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No specific software was used to collect data. |
| Data analysis | We use python python (https://www.python.org) and pytorch (https://pytorch.org/) to analysis the data. The analysis code can be found at https://github.com/peterhan91/CLEAR |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this study are publicly available: The MIMIC-CXR dataset is publicly available from PhysioNet at https://physionet.org/content/mimic-cxr/2.1.0/. The CheXpert-Plus dataset can be accessed at https://aimi.stanford.edu/datasets/chexpert-plus. The ReXGradient data can be accessed at https://huggingface.co/datasets/rajpurkarlab/ReXGradient-160K. The PadChest dataset is available from https://bimcv.cipf.es/bimcv-projects/padchest. The VinDr-CXR dataset can be

obtained from https://vindr.ai/cxr. The Indiana University Chest X-ray Collection is accessible through https://openi.nlm.nih.gov/. The CheXpert test set requires official competition registration at https://github.com/rajpurkarlab/cheXpert-test-set-labels.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | We did not perform sex and gender analysis. |
| Reporting on race, ethnicity, or other socially relevant groupings | We did not collect or use any covariates regarding race, ethnicity, and other social groupings at any stage of the study. |
| Population characteristics | We report dataset sizes (e.g., number of patients and images) but did not collect or use covariates pertaining to population-level demographic characteristics at any stage of the study. |
| Recruitment | No patient recruitment was necessary in this study. |
| Ethics oversight | This study utilized only publicly available, de-identified datasets that have received appropriate institutional review board approvals from their respective institutions. No additional patient data were collected for this research.<br>The MIMIC-CXR, CheXpert-Plus, ReXGradient, PadChest, VinDr-CXR, and Indiana University datasets were all previously de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision.<br>As this study involved only secondary analysis of existing, publicly available data, no additional ethics approval was required. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We leveraged the full available size of each dataset. Pretraining used 873,342 image–text pairs (239,091 patients) aggregated from MIMIC-CXR, CheXpert-Plus, and ReXGradient. Supervised experiments used 368,960 images from MIMIC-CXR (64,586 patients) and were evaluated on the 500-patient CheXpert test set. External validation was performed on VinDr-CXR (3,000 images), PadChest (7,943 images), and Indiana CXR (7,466 images). |
| Data exclusions | To ensure the quality of collected medical data, we excluded data samples with incomplete images or missing labels. |
| Replication | We confirmed reproducibility of our supervised classification experiments by repeating all training with 20 different random seeds and reporting mean and standard deviation of performance metrics. We also performed external validation on three independent datasets (VinDr-CXR, PadChest, Indiana CXR) to verify the generalizability of results. |
| Randomization | Not applicable. This was a secondary analysis of existing observational datasets, with no random allocation of patients or images. For supervised experiments, train/validation/test splits followed established dataset protocols. |
| Blinding | During the data processing, all data were first de-identified to remove any patient related information. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Plants

| | |
|---|---|
| Seed stocks | N/A |
| Novel plant genotypes | N/A |
| Authentication | N/A |