nature portfolio

Corresponding author(s): Ryutaro Tanno, David Barrett

Last updated by author(s): 20/Aug/2024

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 *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 *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	We implemented the annotation tool for conducting the expert evaluation using an internal software.
Data analysis	We used Python and standard visualisation/analysis toolkits (e.g., matplotlib and scipy) for data analysis. We also used CheXPert labeller, an open-sourced NLP software for pre-processing radiology reports.

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

Below is the data availability statement we included in our manuscript:
"MIMIC-CXR27, one of the real-world datasets used in the development of Flamingo-CXR is accessible by researchers and can be downloaded from https://physionet.org/content/mimic-cxr upon completion of the required training. IND1, the other de-identified chest X-ray dataset used in this study cannot be made

publicly available because the authors do not have the rights to do so. Interested researchers should contact info@apollohospitals.com to inquire about access to the IND1 dataset; requests will be subject to Apollo's consideration and applicable ethical and legal requirements. Further enquiries about the data used in this study may be addressed to the corresponding authors with a maximum response time of two weeks. "

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This information has not been collected.
Population characteristics	We evaluate our AI system with a group of radiologists located in two different countries, namely India and the US.
Recruitment	We ensured that the recruited radiologists are capable for performing the evaluation task reliably through a series of written/verbal trainings and by making sure that they hold board certifications of the countries of their residence. We recruited experts from India and the USA to ensure diversity of raters.
Ethics oversight	The use of de-identified retrospective datasets was reviewed by Advarra IRB (Columbia, MD), which determined that it was exempt from further review under 45 CFR 46. The involvement of clinicians in this study, using the same de-identified retrospective data is also covered in this waiver.

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	Sample size was determined by the maximum number of reports and replications viable a budget constraint, while making sure that the sample size (550 reports x 4 annotations per example) in this study exceeds the relevant prior publications such as [1] where 500 reports with 3 annotations per report are used, and [2] where 250 reports x 1 annotation per report are used: [1] J. Huang, L. Neill, M. Wittbrodt, D. Melnick, M. Klug, M. Thompson, J. Bailitz, T. Loftus, S. Malik, A. Phull, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Network Open, 6(10):e2336100–e2336100, 2023. [2] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al.Towards generalist biomedical AI.NEJM AI, 1(3):AIoa2300138, 2024a.
Data exclusions	There were a several cases that we excluded from the clinician-AI rater study. Concretely, there were 4 IND1 cases and 7 MIMIC-CXR cases where the radiologists did not complete part (d) of the edit correction task in accordance with the instructions; and these were excluded from the clinician-AI rater study.
Replication	Each evaluation task is conducted by 4 experts (2 based in the US and 2 in India).
Randomization	Both examples / human participants were randomly allocated to the annotation tasks.
Blinding	The investigators were completely blinded to the group allocation

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

Methods

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