nature portfolio

Corresponding author(s): Jayashree Kalpathy-Cramer (Manuscript #21-0496A)

Last updated by author(s): Nov 15, 2021

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	N/A
Data analysis	All code from our method is available at: https://github.com/QTIM-Lab/MedicalDomainExpansion

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

The main data supporting the results in this study are available within the paper and its Supplementary Information. The raw datasets from DMIST, MGH, North America, and Nepal is not publicly available due to patient privacy restrictions, though potential collaborators are directed to the corresponding author.

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	No sample size calculations were performed. The DMIST, ICROP, and Nepal datasets were shown to be sufficient size for training deep learning models from prior published studies (https://www.sciencedirect.com/science/article/abs/pii/S1546144020305391, https://jamanetwork.com/journals/jamaophthalmology/article-abstract/2680579, https://www.sciencedirect.com/science/article/abs/pii/S2468653020305005?via%3Dihub). The MGH dataset was never used in a prior study, but is shown empirically to be of sufficient size for training deep learning models in this study.
Data exclusions	The DMIST dataset contained 4 data formats: 12-bit monochrome 1 (30.3%), 12-bit monochrome 2 (11.2%), 14-bit monochrome 1 (58.0%), and 14-bit monochrome 2 (0.5%). The 14-bit monochrome 2 images were excluded to ensure that each image data format included in our study had adequate representation for training of our deep learning model. For MGH, we obtained digital screening mammograms from Massachusetts General Hospital (MGH) from 2010. Patients who had prior surgery or implants were excluded. For the North American and Nepal datasets, images were excluded if treatment (such as laser photocoagulation) was visible, if camera artifact obscured >50% of the image, or if retinal detachment was present (stage ≥4).
Replication	Each model was replicated 5 times.
Randomization	All datasets were randomly divided into training, validation, and test sets on the patient level.
Blinding	Blinding is not relevant to this study as there were no human evaluation of the different experimental groups.

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
☒	☐ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The DMIST data is from a previously completed clinical trial: https://clinicaltrials.gov/ct2/show/NCT00008346
Study protocol	Study protocol for DMIST was previously described: https://www.nejm.org/doi/full/10.1056/nejmoa052911
Data collection	Data collection for DMIST dataset was described previously: https://www.nejm.org/doi/full/10.1056/nejmoa052911 Dataset collection for the MGH dataset is described in the paper. Data collection for the ICROP dataset was described previously: https://jamanetwork.com/journals/jamaophthalmology/article-abstract/2680579 Data collection for the Nepal dataset was described previously: https://www.sciencedirect.com/science/article/abs/pii/S2468653020305005?via%3Dihub
Outcomes	For the mammography datasets, breast density was determined by clinical radiologists

Outcomes	For ROP datasets, stage was determined by clinical ophthalmologists