nature portfolio

Corresponding author(s): Muhammad Dawood

Last updated by author(s): Jan 14, 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 *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

• Whole slide images (WSIs) of TCGA patients used in the study can be downloaded from the NIH Genomic Data Common Portal at this link: https://portal.gdc.cancer.gov/.
• The genomic data and clinical data of patients in TCGA, METABRIC, COAD-DFCI, MSK-LUAD and CPTAC cohorts can be downloaded from cBioPortal https://www.cbioportal.org/.
• ABCTB data and images were obtained from the Australian Breast Cancer Tissue Bank. The corresponding author could be contacted to facilitate access to ABCTB data.

Data analysis

tiatoolbox (1.3.0) - https://github.com/TissueImageAnalytics/tiatoolbox/
torch geometric (2.1.0.post1) - https://pytorch-geometric.readthedocs.io/en/latest/install/installation.html
PyTorch (1.12.1+cu102) - https://pytorch.org/ PRID: SCR_018536
Scipy (1.7.3) - http://www.scipy.org/ PRID: SCR_008058
NumPy (1.22.4) - http://www.numpy.org/ PRID: SCR_008633
Pandas (1.5.3) - https://pandas.pydata.org/ PRID: SCR_018214
Matplotlib (3.3.4) - https://matplotlib.org/ PRID: SCR_008624
OpenCV (4.5.5) - https://opencv.org/
statsmodels (0.12.2) - https://www.statsmodels.org/ PRID: SCR_016074

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 data used in this study—including histological and molecular information for TCGA and CPTAC cohorts are primarily obtained from the cBioPortal. Detailed manifest files containing unique TCGA or CPTAC identifiers for each sample are provided in our Project GitHub repository: https://github.com/engrodawood/HistBiases. Additional information on data access and usage can be found in the repository documentation.

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	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Detailed population characteristics (e.g., genotypic information, age, etc.) for the samples analyzed in the study can be accessed through cBioportal.
Recruitment	We do not recruit any participants. Instead, we retrospectively analyzed data from 8,221 patients with breast, colorectal, endometrial, and lung cancers across four cohorts: The Cancer Genome Atlas (TCGA), the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), Memorial Sloan Kettering Cancer Centre (MSK), and Dana-Farber Cancer Institute (DFCI).
Ethics oversight	Biomedical and Scientific Research Ethics Committee (BSREC) University of Warwick Approved the study under application ID BSREC 16/21-22.

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 analyzed samples from existing datasets and demonstrated the validity of our findings across a large sample size of n = 8,221, derived from multiple datasets
Data exclusions	We excluded Whole Slide Images with missing meta data. The exclusion criteria is clearly explained in the manuscript.
Replication	The results reported in the manuscript can be easily reproduced using the project's GitHub repository: https://github.com/imuhdawood/HistBiases. The repository includes source code for codependency analysis, machine learning models, as well as permutation testing and stratification analysis, along with detailed documentation. These resources can be used to explore potential biases in machine learning models. Additionally, predictions from the models on both training and test datasets are provided, enabling downstream bias analysis without the need to rerun the entire pipeline for each model.
Randomization	Randomization of sample for cross-validation folds (4-folds) was performed at random with stratification. The model is trained on 3 folds and the performance is evaluated on the held out fold.
Blinding	Data of all patients were analyzed anonymously. Not additional blinding was done.

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

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	Not applicable
Study protocol	Not applicable
Data collection	Not applicable
Outcomes	Not applicable

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable