scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Multi-Stain Breast Cancer Histological Whole-Slide-Image Data Set from Routine Diagnostics

Philippe Weitz[1,11]✉, Masi Valkonen[2,11], Leslie Solorzano[1,11], Circe Carr[2], Kimmo Kartasalo[1], Constance Boissin[1], Sonja Koivukoski[3], Aino Kuusela[2], Dusan Rasic[4,5], Yanbo Feng[1], Sandra Sinius Pouplier[4,10], Abhinav Sharma[1], Kajsa Ledesma Eriksson[1], Leena Latonen[3,6], Anne-Vibeke Laenkholm[4,10], Johan Hartman[7,8,11], Pekka Ruusuvuori[2,9,11] & Mattias Rantalainen[1,8,11]✉

The analysis of FFPE tissue sections stained with haematoxylin and eosin (H&E) or immunohistochemistry (IHC) is essential for the pathologic assessment of surgically resected breast cancer specimens. IHC staining has been broadly adopted into diagnostic guidelines and routine workflows to assess the status of several established biomarkers, including ER, PGR, HER2 and KI67. Biomarker assessment can also be facilitated by computational pathology image analysis methods, which have made numerous substantial advances recently, often based on publicly available whole slide image (WSI) data sets. However, the field is still considerably limited by the sparsity of public data sets. In particular, there are no large, high quality publicly available data sets with WSIs of matching IHC and H&E-stained tissue sections from the same tumour. Here, we publish the currently largest publicly available data set of WSIs of tissue sections from surgical resection specimens from female primary breast cancer patients with matched WSIs of corresponding H&E and IHC-stained tissue, consisting of 4,212 WSIs from 1,153 patients.

Background & Summary

Breast cancer is the most common cancer in women globally with 11.7% of all cases and the fourth most common cause of cancer deaths in women with 6.9% of all cancer deaths[1]. Biomarker assessment through IHC staining, particularly of the hormone receptors ER and PGR, for oestrogen and progesterone, respectively, as well as for the receptor for human epidermal growth factor 2, HER2, has become an essential component of the routine pathology workflow where available[2,3]. Another biomarker that is routinely assessed through IHC staining in some countries is KI67. The *International KI67 in Breast Cancer Working Group* (IKWG) currently recommends KI67 scoring at least in patients classified as ER-positive and HER2-negative based on IHC scores[4].

Automated IHC biomarker scoring with image analysis software can enhance its validity and reproducibility. The IKWG found that e.g. for KI67 scoring, automated scoring with QuPath[5] shows outstanding reproducibility[6]. In recent years, image analysis in the context of computational pathology has advanced for a whole range of applications. This has at least in part been facilitated by large, publicly available WSI data sets, such as the resources provided by the TCGA research network. Publicly available data does not only provide development data to the research community, but perhaps even more importantly, it also allows for a comparable benchmarking of novel methods on the same test data.

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [2]Institute of Biomedicine, University of Turku, Turku, Finland. [3]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [4]Department of Surgical Pathology, Zealand University Hospital, Roskilde, Denmark. [5]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [6]Foundation for the Finnish Cancer Institute, Helsinki, Finland. [7]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden. [8]MedTechLabs, BioClinicum, Karolinska University Hospital, Stockholm, Sweden. [9]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [10]Present address: Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [11]These authors contributed equally: Philippe Weitz, Masi Valkonen, Leslie Solorzano, Johan Hartman, Pekka Ruusuvuori, Mattias Rantalainen. ✉e-mail: philippe.weitz@ki.se; mattias.rantalainen@ki.se

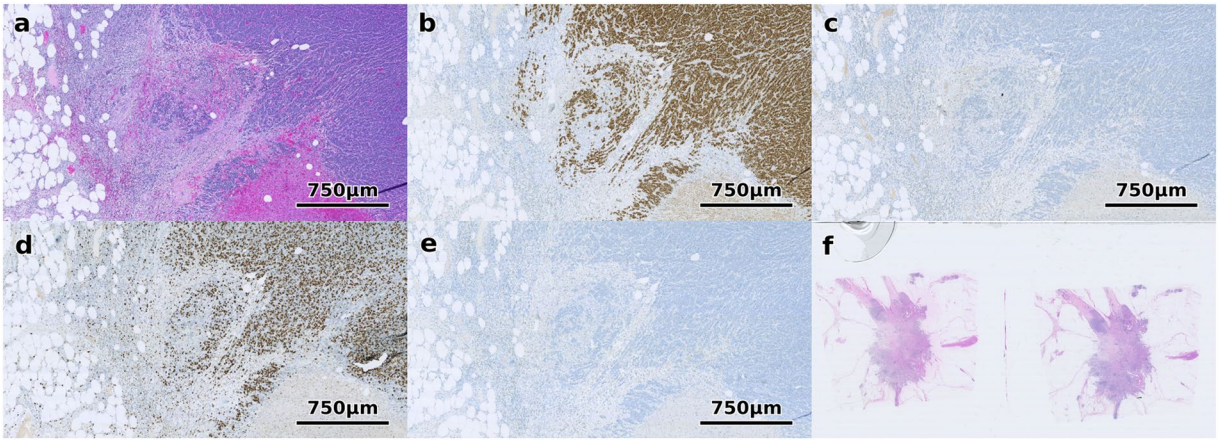

Fig. 1 Example of an H&E-stained tissue region with corresponding IHC-stained tissue with all four routine diagnostic stains available in this data set. (**a**) shows H&E, (**b**) ER, (**c**) HER2, (**d**) Ki67 and (**e**) PGR. (**f**) shows an example of a WSI that was excluded since it contains multiple tissue sections.

There are several application areas where multi-modal (e.g. multiple stains) WSI image data is required. This includes development of high performing WSI registration (the spatial alignment of corresponding tissue in two or more WSIs) methods, which is an enabling technology both for research and diagnostics. It can allow clinicians to fuse information from different IHC-stains in WSI viewers. Combining information from H&E-stained tissue with corresponding IHC-stained tissue regions can e.g. be of critical importance when investigating resection borders with respect to malignancy, which can be very time consuming without tissue alignment. Some commercial IHC scoring softwares also align H&E and IHC-stained tissue to enable pathologists to contextualise automated scoring results. In research projects, WSI registration can facilitate stain-guided learning[7–9], virtual staining[10–12], 3D reconstruction[13,14] and the transfer of annotations between different WSIs and stains. However, there is currently a lack of publicly available data sets that include WSIs from H&E-stained tissue sections with matched IHC-stained tissue from the same tumour, despite the importance of IHC for pathological diagnosis.

To promote and enable further research in this domain, we have published the ACROBAT (AutomatiC Registration Of Breast cAncer Tissue) data set[15], which consists of 4,212 WSIs from 1,153 female primary breast cancer patients. For each patient, the data set[15] contains one WSI of H&E stained tissue and up to four WSIs with tissue that was stained with the routine diagnostic IHC markers ER, PGR, HER2 and KI67. An example of a case from the data set[15] with all four IHC antibodies available is depicted in Fig. 1.

The data set[15] was initially collected as part of the CHIME study (chimestudy.se) at Karolinska Institutet (Stockholm, Sweden). The primary purpose of the CHIME study is to advance precision medicine through computational pathology, based on population representative patient cohorts. Histopathology slides came from the routine clinical diagnostic workflow, with WSIs generated using high-throughput histopathology slide scanners at Karolinska Institutet.

The primary purpose of publishing this data set[15] was to enable the ACROBAT WSI registration competition (acrobat.grand-challenge.org), which took place in the scope of the MICCAI (Medical Image Computing and Computer Assisted Intervention) 2022 conference.

While the primary purpose of this data set[15] is the development of WSI registration methods, we believe that there could also be other use cases for the data. These may include the development of digital staining and stain transfer methods, as well as novel methods in stain-guided learning. Further applications may include the development of tissue segmentation and classification algorithms, the development of artefact detection or correction methods and unsupervised pre-training e.g. of convolutional neural networks (CNNs), which are then fine-tuned for specific tasks. We therefore hope that this data set[15] can contribute towards the advance of WSI registration methods but also other research activities in the domain of computational pathology.

Methods

Data selection and splits for ACROBAT data set. The CHIME breast cancer study is based upon a retrospective cohort study design. Consecutive female breast cancer cases diagnosed between 2012 and 2018 at Södersjukhuset (Stockholm, Sweden) were included. The study has approval by the regional ethics review board (Etiksprövningsmyndigheten, Stockholm, Sweden, ref. 2017/2106-31 and amendments 2018/1462-32, 2019-02336). Due to the retrospective nature of the study, consent was not required. The archived histopathology slides were retrieved and scanned. The training and validation set are a randomly selected subset of the data generated in terms of the CHIME study. The test set WSIs were chosen as a subset of the CHIME study data that has previously been reviewed by a pathologist specialising in breast pathology in the context of another research project. Cases were excluded for the ACROBAT data set[15] only if one of the WSIs contains multiple sections of the same tissue, which occurs in approximately 1% of H&E WSIs in this data set[15], as depicted in Fig. 1f. These cases were excluded since it would be unclear to which of the multiple sections tissue from the corresponding IHC WSI should be aligned to. In this case, a new case was randomly selected and included after the corresponding quality control for multiple tissue sections.

	Training	Validation	Test	Total
Cases	750 (65.0%)	100 (8.7%)	303 (26.3%)	1153
WSIs	3406 (80.9%)	200 (4.7%)	606 (14.8%)	4212
Stain and antibodies				
H&E	750 (22.0%)	100 (50.0%)	303 (50.0%)	1153 (27.4%)
ER	732 (21.5%)	29 (14.5%)	84 (13.9%)	845 (20.1%)
KI67	732 (21.5%)	29 (14.5%)	82 (13.5%)	843 (20.0%)
PGR	728 (21.4%)	28 (14.0%)	81 (13.4%)	837 (19.9%)
HER2	464 (13.6%)	14 (7.0%)	56 (9.2%)	534 (12.7%)
Scanners				
NanoZoomer S360	559 (16.4%)	38 (19.0%)	205 (33.8%)	802 (19.0%)
NanoZoomer XR (1)	884 (26.0%)	46 (23.0%)	203 (33.5%)	1133 (26.9%)
NanoZoomer XR (2)	1963 (57.6%)	116 (58.0%)	198 (32.7%)	2277 (54.1%)

Table 1. Distribution of cases and WSIs to training, validation and test set, as well as respective distributions of stains, IHC antibodies and scanner models. Three Hamamatsu NanoZoomer scanners were used for WSI digitisation, one NanoZoomer S360 and two NanoZoomer XR.

The training set consists of 3,406 WSIs from 750 patients. Each patient has one H&E WSI and up to four associated IHC WSIs from the routine diagnostic IHC antibodies ER, PGR, HER2, KI67, as depicted in Fig. 1. The validation data set consists of 200 WSIs from 100 patients and the test set consists of 606 WSIs from 303 patients. Each case in the validation and test sets consists of one H&E WSI per case and one IHC WSI, which was randomly selected stratifying for IHC antibody. The test set was furthermore selected by stratifying for clinical covariates by balanced sampling from the three different WSI scanners. Table 1 indicates the distributions of scanners and IHC antibodies in the respective subsets.

Whole slide image scanning. WSIs in the CHIME study that were available at the time of data selection were generated from archived histopathology slides with three Hamamatsu WSI scanners, consisting of one NanoZoomer S360 and two NanoZoomer XRs. Slides were digitised by a trained scanning technician using an automated scanning workflow, with manual rescanning of slides where automated focusing was not successful. Slides were scanned at a resolution of approximately 0.23 µm/pixel and a JPEG compression quality level of 80.

Image processing. The 40X NDPI WSIs were first anonymized and then converted to pyramidal TIFF files with 10X and lower resolutions. Macro images and other identifying information in the WSI metadata were removed with code available from github.com/bgilbert/anonymize-slide. Then, file names were generated consisting of a random case ID, the stain or antibody of the WSI, as well as the name of the respective set out of training, validation or test. TIFF files were then extracted using the libvips[16] command *im_vips2tiff* at 10X and lower magnifications, with 7 to 9 magnification levels depending on the WSI available and a downsampling factor of 2 between these levels. This reduces the storage requirements of the data set[15] from 10.13 TB to 482 GB, likely without impacting the performance of image registration algorithms as registration is typically performed at fairly low resolutions with diminishing to no improvements at higher resolutions[17].

Annotation workflow. Members of the ABCAP research consortium (abcap.org) were enrolled as annotators to generate landmarks, including 13 individuals in total. All annotators have previous experience from working directly with WSIs in a research context and have received corresponding training. Two of the annotators have pathologist training. Landmark annotations were generated using a customised version of TissUUmaps[18]. Image pairs in the validation data were annotated by one annotator, whereas each image pair in the test data was annotated by two annotators in two annotation phases. No annotations were generated for the training data. All annotations were conducted with the original NDPI files at 40X magnification. Counting landmarks from both annotation phases independently, annotators generated 35,760 landmark pairs in total.

In the first phase of the annotation process, which is the same for the validation and test data, annotators were shown one H&E and one IHC stained section side-by-side and were asked to mark 50 corresponding landmarks in both images, inserting first the IHC point and then the H&E point. In order to ensure the precision of annotations, annotators were asked to zoom the field of view to a level such that the TissUUmaps scale bar indicated 250 µm or less.

In the second phase, which was only applied to the test data, annotators were provided with modified annotation files from the first round of annotations. Landmark coordinates in the IHC image were fixed in place, while for the H&E random uniform noise of $[-500, 500)$ pixels ($\pm 115\,\mu m$) was added to both the X and Y coordinates. Annotators were then asked to move the H&E landmarks to match the corresponding ones in the IHC WSI. Annotators were chosen randomly such that phase one and phase two annotations were created by different observers for each WSI. Detailed annotation guidelines for both phases are available from github.com/rantalainenGroup/ACROBAT.

Column name	Definition
anon_id	Anonymized case id. The same case ids can occur in the training, validation and test set, so combine this with the set column to uniquely identify a patient.
set	Indicates which data split this file belongs to.
stain	Indicates the stain for WSIs of H&E stained tissue or IHC antibody for WSIs of IHC stained tissue.
filename	Indicates the filename that the meta data refers to.
vendor	Vendor of the scanner used to generate the WSI. Always Hamamatsu in this dataset.
model	Indicates the Hamamatsu scanner model used to generate the respective WSI.
mpp_lvl_X	Indicates the microns-per-pixel for pyramidal level X, where X is in [0, 8]. Not all WSIs have 9 levels, the value if a level does not exist is NaN.
magnification_lvl_X	Indicates the magnification for pyramidal level X, where X is in [0, 8], starting at 10X. Not all WSIs have 9 levels, the value if a level does not exist is NaN.

Table 2. Overview of WSI-wise information that is available for download alongside the ZIP archives that contain the WSIs as a CSV file.

Data Records

We published 4,212 WSIs of breast cancer resection specimens stained with H&E or IHC (ER, PGR, HER2, KI67) originating from 1,153 patients on the Swedish National Data Service SND[15]. All WSIs are provided as pyramidal TIFF files, starting at 10X resolution (ca. 0.92 µm/pixel) and lower resolutions. The naming convention of all WSIs follows the pattern *caseid_stain_set.tiff* where *caseid* indicates a randomly generated case ID, *stain* either H&E or the IHC antibody used, and *set* whether the file belongs to *train*, *valid* or *test*. Furthermore, there is a CSV table that indicates the microns-per-pixel at the first level in the respective TIFF files, the stain, the IHC antibody and the data split for each file name. This table is summarised in Table 1.

The data set[15] is available for download in seven separate ZIP archives, five for the training data (train_part1.zip (71.47 GB), train_part2.zip (70.59 GB), train_part3.zip (75.91 GB), train_part4.zip (71.63 GB) and train_part5.zip (69.09 GB)), one for the validation data (valid.zip 21.79 GB) and one for the test data (test.zip 68.11 GB). Lists of files and checksums in SHA1 format are available in order to ascertain completeness and integrity of the data set[15] after download. Furthermore, a CSV file (df_acrobat_meta.csv) with a table with additional information for each WSI is available, alongside a ReadMe file (df_acrobat_meta_readme.txt) that explains the columns in this table. Table 2 contains these explanations. No additional clinical information beyond these columns is provided.

Technical Validation

All WSIs originate from slides that were used in the routine diagnostic workflow. The tissue samples have therefore each been reviewed by at least one specialty pathologist using a microscope during the initial diagnosis. The macro images of all WSIs in the data set[15] were reviewed by at least one observer in order to exclude WSIs with multiple tissue sections of the same resection specimen and in order to confirm that H&E and IHC tissue sections show corresponding tissue. All WSIs in the validation set were reviewed by at least one and all WSIs in the test set were reviewed by at least two human annotators during the landmark generation at 40X resolution. All WSIs in the test data were furthermore reviewed by a specialty pathologist while generating annotations for a research project that is independent from the ACROBAT challenge at 40X resolution, confirming their usability. The majority of WSIs in this study have also been used in other research studies, which further supports the validity of the data set[15,19]. There are several studies that use WSIs that were generated using the same scanners and workflow[20–22]. Some of the WSIs included in the data set[15] contain artefacts. These WSIs were deliberately left in the data set[15], in order to be able to assess the robustness of suggested registration methods.

The quality of landmarks in the test data can be assessed by computing the distances between the two human annotators. Landmarks with a distance between annotators of more than 115 µm were excluded, which was chosen as a threshold in correspondence to the noise added for the second annotation phase.

Usage Notes

Pyramidal TIFF files are compatible with OpenSlide[23] and can e.g. be inspected with QuPath. Registration algorithms typically align WSIs iteratively starting at low resolutions. With OpenSlide, lower resolution versions of the WSI can be obtained through the different levels of the TIFF files, which makes additional computations for downsampling obsolete. ACROBAT Github repository at github.com/rantalainenGroup/ACROBAT provides code to visually inspect landmarks, either only in IHC or also paired H&E landmark after registration by the user.

The data set[15] is split into a training, a validation, and a test set. To evaluate the performance of registration methods, landmarks for the IHC WSIs in the validation and test data, are released publicly, whereas the target H&E landmarks can be used to quantify registration performance through an automated evaluation tool. Registered validation set landmarks can be submitted at acrobat.grand-challenge.org to receive performance metrics based on these landmarks.

Code availability

NDPI files were anonymized with code available from github.com/bgilbert/anonymize-slide. The libvips package for pyramidal TIFF extraction is available from github.com/libvips/libvips. The tool used to generate landmark annotations is based on TissUUmaps, which is available from tissuumaps.github.io. Code used for displaying landmarks in a surrounding tissue region, code for computing registration performance metrics, as well as the annotator protocols are available from github.com/rantalainenGroup/ACROBAT.

Received: 16 January 2023; Accepted: 27 July 2023;
Published online: 24 August 2023

References

1. Sung, H. *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Allison, K. H. *et al*. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. *J. Clin. Oncol.* **38**, 1346–1366 (2020).
3. Wolff, A. C. *et al*. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* **31**, 3997–4013 (2013).
4. Nielsen, T. O. *et al*. Assessment of Ki67 in Breast Cancer: Updated Recommendations from the International Ki67 in Breast Cancer Working Group. *J. Natl. Cancer Inst.* **113**, 808–819 (2021).
5. Bankhead, P. *et al*. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
6. Acs, B. *et al*. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab. Invest.* **99**, 107–117 (2019).
7. Su, A. *et al*. A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images. *NPJ Precis Oncol* **6**, 14 (2022).
8. Turkki, R., Linder, N., Kovanen, P. E., Pellinen, T. & Lundin, J. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J. Pathol. Inform.* **7**, 38 (2016).
9. Valkonen, M. *et al*. Cytokeratin-Supervised Deep Learning for Automatic Recognition of Epithelial Cells in Breast Cancers Stained for ER, PR, and Ki-67. *IEEE Trans. Med. Imaging* **39**, 534–542 (2020).
10. Burlingame, E. A. *et al*. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Sci. Rep.* **10**, 17507 (2020).
11. Wieslander, H., Gupta, A., Bergman, E., Hallström, E. & Harrison, P. J. Learning to see colours: Biologically relevant virtual staining for adipocyte cell images. *PLoS One* **16**, e0258546 (2021).
12. de Haan, K. *et al*. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
13. Kartasalo, K. *et al*. Comparative analysis of tissue reconstruction algorithms for 3D histology. *Bioinformatics* **34**, 3013–3021 (2018).
14. Song, Y., Treanor, D., Bulpitt, A. J. & Magee, D. R. 3D reconstruction of multiple stained histology images. *J. Pathol. Inform.* **4**, S7 (2013).
15. Rantalainen, M. & Hartman, J. ACROBAT - a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology. *Swedish National Data Service (SND)* https://doi.org/10.48723/w728-p041 (2023).
16. Martinez, K. & Cupitt, J. VIPS - a highly tuned image processing software architecture. *IEEE International Conference on Image Processing 2005* **2**, II–574 (2005).
17. Lotz, J., Weiss, N., van der Laak, J. & Heldmann, S. Comparison of Consecutive and Re-stained Sections for Image Registration in Histopathology. Preprint at https://arxiv.org/abs/2106.13150 (2022).
18. Solorzano, L., Partel, G. & Wählby, C. TissUUmaps: interactive visualization of large-scale spatial gene expression and tissue morphology data. *Bioinformatics* **36**, 4363–4365 (2020).
19. Wang, Y. *et al*. Improved breast cancer histological grading using deep learning. *Ann. Oncol.* **33**, 89–98 (2022).
20. Wang, Y. *et al*. Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression-Morphology Analysis in Breast Cancer. *Cancer Res.* **81**, 5115–5126 (2021).
21. Weitz, P., Wang, Y., Hartman, J. & Rantalainen, M. An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. https://doi.org/10.1109/iccvw54120.2021.00074 (IEEE, 2021).
22. Liu, B. *et al*. Using deep learning to detect patients at risk for prostate cancer despite benign biopsies. *iScience* **25**, 104663 (2022).
23. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).

Acknowledgements

We acknowledge support from Stratipath and Karolinska Institutet sponsoring the ACROBAT challenge prize; MICCAI society for hosting the ACROBAT challenge, and Nguyen Thuy Duong Tran for support with digitising histopathology slides. We acknowledge funding from: Vetenskapsrådet (Swedish Research Council), Cancerfonden (Swedish Cancer Society), ERA PerMed (ERAPERMED2019-224-ABCAP), MedTechLabs, Swedish e-science Research Centre (SeRC), VINNOVA, SweLife, Academy of Finland (#341967, #334782, #335976, #334774), Cancer Foundation Finland, University of Turku Graduate School, Turku University Foundation, Oskar Huttunen Foundation, David and Astrid Hägelén Foundation.

Author contributions

P.W., M.V., L.S. jointly organised the ACROBAT challenge. M.R., P.R. jointly supervised the ACROBAT challenge organisation. P.W., M.R., M.V. and P.R. jointly conceptualised the ACROBAT challenge. P.W. selected and verified the data set. P.W. drafted the manuscript. P.W., K.K., M.V. processed the images. L.S., M.V. implemented the annotation infrastructure. M.V., L.S., P.W. generated the annotation instructions. C.C., C.B., S.K., A.K., D.R., Y.F., S.P., P.W., M.V., L.S., K.K., A.S., K.L.E. (in order of contribution) generated the landmark annotations. M.R., J.H., P.R., A.L., L.L. acquired funding for this project. All authors contributed to editing the manuscript.

Funding

Open access funding provided by Karolinska Institute.

Competing interests

M.R., J.H. are co-founders and shareholders of Stratipath.

Additional information

Correspondence and requests for materials should be addressed to P.W. or M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023