scientific data

OPEN

DATA DESCRIPTOR

A clinical prostate biopsy dataset with undetected cancer

Eduard Chelebian[1] ✉, Christophe Avenel[1], Helena Järemo[2], Pernilla Andersson[2], Carolina Wählby[1] & Anders Bergh[2]

Prostate cancer is a heterogeneous disease showing variability both among individuals and within a patient. While most cases are indolent, aggressive tumors require early intervention. Accurately predicting tumor behavior is challenging, contributing to overdiagnosis but also undertreatment. Current imaging methods may miss the most malignant areas, leading to biopsies often capturing non-malignant prostate tissue even if cancer is present elsewhere in the organ. This non-malignant tissue, however, holds potential as a source for novel diagnostic and prognostic markers. Our clinical dataset comprises men with raised prostate-specific antigen but whose initial prostate needle biopsies only contained benign tissue. Half of the paired patients remained cancer-free for over eight years, while the others were diagnosed with prostate cancer within 30 months of follow-up. We share these initial benign biopsies to enable the exploration of morphological changes in non-malignant tissue and the potential for improved diagnostic accuracy in the early identification of patients with prostate cancer.

Background & Summary

Prostate cancer (PCa) is one of the most prevalent cancers globally, constituting approximately 7.3% of all cancer cases and 3.8% of all cancer-related deaths, with notable variations across different geographical regions[1]. PCa is characterized by its multifocal and heterogeneous nature, where patients may harbor multiple distinct tumor foci simultaneously, many of which are indolent and not clinically significant[2]. Presently, the primary approach for early PCa detection relies on the measurement of the prostate-specific antigen (PSA) in blood samples. Raised values (> 3 ng/ml) suggest prostate pathology, though not necessarily cancer. The higher the PSA the higher the risk that a clinically significant PCa is present[3]. While the introduction of the PSA test led to higher detection rates of PCa, it has also resulted in a surge in overdiagnosis and overtreatment, given the high prevalence of patients with indolent prostate tumors who do not require treatment[4].

Patients presenting elevated PSA levels generally undergo prostate magnetic resonance imaging (MRI) followed by MRI-guided and/or systematic needle biopsies[5,6]. Biopsies are stained employing hematoxylin-eosin (H&E) staining and examined by microscopy[7]. The International Society of Urological Pathology (ISUP) introduced the ISUP grading system in 2019, building upon the classical microscopy-based Gleason grading system and comprising five distinct grading groups[8]. ISUP grade 1 denotes a low risk of disease progression and where active treatment is unnecessary[9]. While ISUP grades 2 and 3 indicate intermediate risk, ISUP3 is more aggressive than ISUP2. Grades 4 and 5 are the most malignant PCa variants. Prostate needle biopsies typically involve the collection of approximately 12 samples taken from different predetermined locations and MRI-detected lesions. Given the limited size of needle biopsies, sampling less than 0.1% of the prostate volume, there is a risk that clinically significant tumor areas are missed[10]. Consequently, a substantial portion of these initial biopsies result in benign diagnoses, highlighting the need for improved detection methods[11]. Previous studies have indicated that cancers, to grow and spread, need to influence the benign parts of the tumor-bearing organ, and the nature and magnitude of these modifications are related to tumor aggressiveness[12]. Understanding such changes could provide valuable insights for early diagnosis and detection.

Currently, available prostate datasets primarily focus on comprehensive collections with cancer-containing biopsies designed for cancer detection purposes. In this category, the dataset accompanying the Prostate cANcer graDe Assessment (PANDA) Challenge[13] stands out as the largest, including over 10,000 whole-slide images (WSI) biopsies from six different institutions to facilitate the development of a Gleason grading algorithm. Additional efforts have aimed to emulate the common scenario of having benign biopsies with undetected PCa by sampling benign cores in the proximity of detected cancer biopsies[14]. While this approach allows the analysis

[1]Department of Information Technology and SciLifeLab, Uppsala University, 752 37, Uppsala, Sweden. [2]Department of Medical Biosciences, Pathology, Umeå University, 901 85, Umeå, Sweden. ✉e-mail: eduard.chelebian@it.uu.se

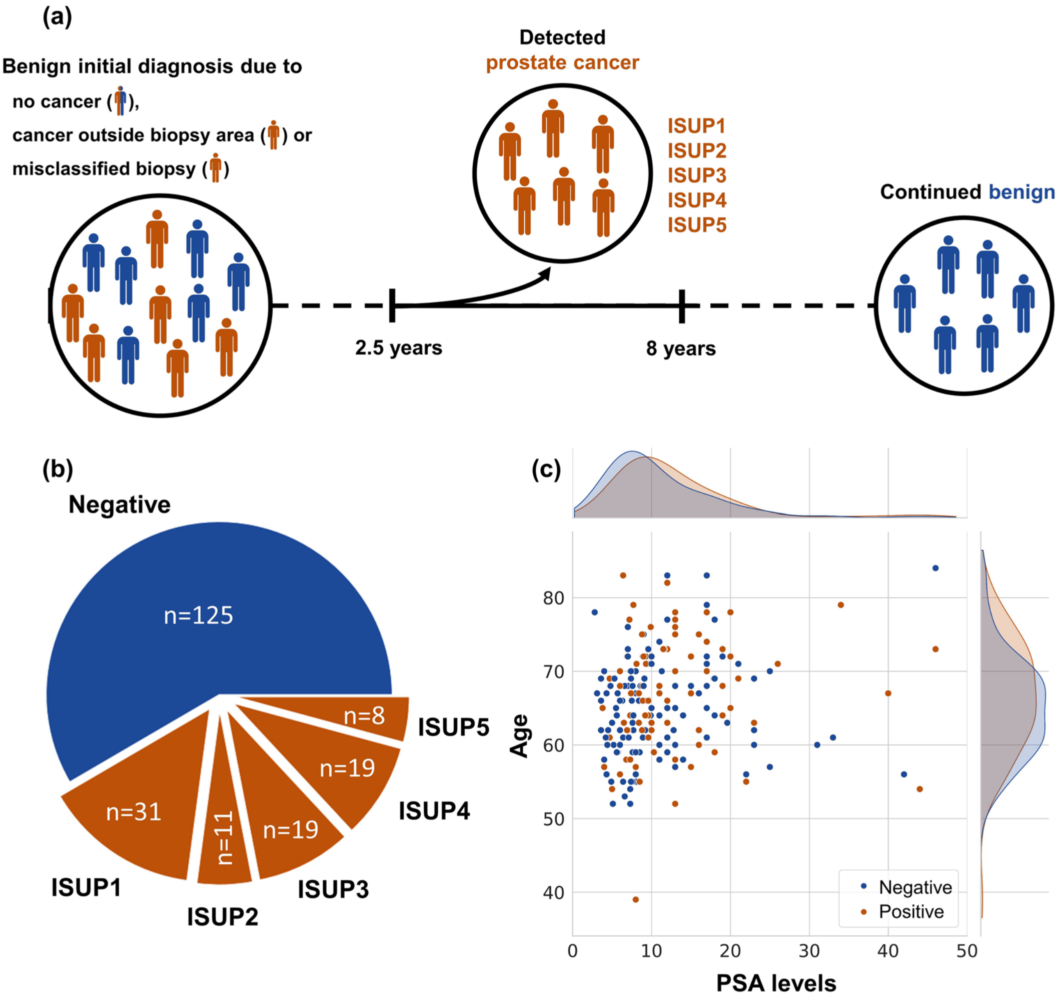

Fig. 1 Summary of the dataset. **(a)** Data collection protocol. **(b)** Distribution of the patients. **(c)** The cohort was selected to have similar PSA and age distribution of negative and positive patients.

of changes in cancer-associated benign tissue, it may be considered too proximal to the lesion compared to the conditions encountered in routine clinical practice. Complementing these approaches, studies such as the DOCUMENT trial have highlighted the utility of epigenetic assays in identifying missed cancers in initial negative biopsies, emphasizing the potential for systematic integration of such tools in improving cancer detection in follow-up biopsies[15]. Thus, there is potential value in exploring biopsies classified as benign but where later follow-up and re-biopsies revealed that clinically significant cancer was present and presumably missed by the initial biopsies.

To address this, we introduce a novel dataset comprising digitized benign prostate H&E biopsies from men with raised PSA. The biopsies were systematically taken from different locations in the prostate and were not guided by MRI as, given the long follow-up times, the patients preceded the extended use of MRI for guiding biopsies. The dataset is uniquely designed, including paired samples from patients with comparable age and PSA levels, all initially diagnosed as benign but with different outcomes upon subsequent follow-ups and re-biopsies. While some patients remained cancer-free during at least eight years of follow-up, others were diagnosed with PCa within the subsequent 30 months of follow-up. The final processed dataset encompasses 213 patients, resulting in a total of 587 H&E prostate needle biopsies. Among these, 125 control patients with 333 biopsies exhibited no cancer development in at least eight years following the initial diagnosis. Conversely, 88 case patients with 254 biopsies were diagnosed with PCa of various ISUP grades within the 30 months following the initial diagnosis. Each case patient is accompanied by up to three control patients, paired for similar age, PSA and year of diagnosis, ensuring a robust comparative analysis. In Fig. 1, we provide a schematic view of the dataset and propose three reasons why an initial biopsy could appear benign: (1) no cancer was present initially, even for cases later diagnosed; (2) cancer was present but missed due to sampling outside the affected area; or (3) cancer was misclassified, though this is unlikely in our dataset, it can be still worth considering.

In this work, we present the dataset used in a prior clinically oriented work[16]. We envision the application of this dataset for uncovering new insights within the initially benign biopsies. Our scope goes beyond traditional diagnosis, aiming to study the long-range morphological effects induced by PCa in the benign parts of the organ. The dataset offers a unique opportunity to understand PCa progression, enabling the identification of potential

Column name	Explanation
Patient_N	Anonymous patient ID.
PSA	Prostate-specific antigen value at the time of biopsy collection.
Age	Age range of the patient at the time of biopsy collection.
Positive	Cancer status in the subsequent 30 months of follow-up after biopsy collection.
ISUP	ISUP grade of the cancer found in the subsequent 30 months of follow-up after biopsy collection.

Table 1. Per slide information included alongside the images.

markers for early detection by leveraging current computational methods. This, in turn, may open avenues for new diagnostic and therapeutic strategies.

Methods

Patient selection.

Patients included in the dataset were retrospectively selected from records at the University Hospital of Umeå (*Norrlands universitetssjukhus*) in Umeå, Sweden between the years 1997 and 2016. The initial step involved identifying patients who had undergone a non-initial positive biopsy. Among this subset, individuals with an initial negative biopsy within the 30 months prior to the detection of PCa were selected. For each patient meeting these criteria, up to three control patients were selected based on similar PSA levels, age, and year of diagnosis. Control patients were required to have had initial negative biopsies and to have continued negative at re-biopsies for at least eight years of follow-up. Note that MRI-guided biopsies were only introduced in the region in 2015 so, due to the long follow-up times needed for this setup, all the patients underwent systematic biopsies. To comply with ethical standards and ensure patient confidentiality, all data was anonymized upon collection. Ethical approval for use and publication of the data was granted by the *Regionala Etikprövningsnämnden* in Umeå, DNR 2010/366-31M. The data remains fully anonymized and does not allow for identification of any individuals.

In the process of patient selection, clinical data was also acquired from the registries at the Department of Pathology at the University Hospital of Umeå. Apart from the selection parameters such as PSA levels and age, a key factor of our dataset is the ISUP grade assigned to patients upon the late discovery of PCa. Less aggressive PCa lesions, classified as ISUP1, are generally associated with fewer cancer-related alterations in adjacent tissues[9,12]. Additionally, ISUP1 lesions can exhibit morphological characteristics that resemble benign tissue, thus potentially posing a challenge in the initial diagnosis of the biopsies. The unique aspect of our dataset lies in the identification of patients initially diagnosed as benign whose subsequent re-biopsies revealed clinically significant cancer (ISUP2–5) within the 30 following months. Studying the morphological nuances between them is fundamental to understanding the progression dynamics of PCa.

Digitization.

The biopsy specimens underwent digitization using the 3D Histech Pannoramic 250 scanner (3D HISTECH Ltd., Budapest, Hungary), operating at a magnification of 20 × with a pixel size of 0.2428 μm/pixel. The process yielded MRXS files in `.mrxs` format, including eight levels of pyramidal downsampling. The pyramidal format enables efficient data storage and manipulation in subsequent computational analyses.

Preprocessing.

To ensure the technical viability of the dataset, we implemented the preprocessing workflow proposed by Lu *et al.*[17] The preprocessing consists of WSI tissue detection and segmentation modules, a fundamental first step in digital pathology. This generates tissue masks, enabling the identification and exclusion of biopsies with insufficient tissue or noticeable artifacts.

Statistical analysis.

In order to confirm the effectiveness of the patient pairing, multivariate logistic regression analysis was conducted to explore the relationship between patient age and PSA at the time of benign diagnosis with the future development of PCa. We explored the R-squared (R^2) of the model as well as the individual p-values for both variables. Additionally, we calculated the area under the receiver operating characteristic curve (AUC) to establish a baseline.

Data Records

The complete clinical prostate biopsy dataset, where the cancer was initially undetected, is openly accessible at the Swedish National Data Service[18]. This dataset includes 587 WSI H&E-stained prostate needle biopsies, originating from 213 patients initially diagnosed as benign.

Each slide in the dataset is provided in MRXS format, which includes a `.mrxs` file with the image and a folder containing `.dat` files. The naming convention for each slide follows the pattern *patient_NNN_SSS*. Here *NNN* represents the anonymous code for the patient and *SSS* indicates the specific biopsies within the slide. For instance, *patient_137* has associated four image files: *patient_137_ABC*, *patient_137_DEF*, *patient_137_GHK* and *patient_137_LMN* with three parallel slides each. On average, each patient has three associated slide files.

Each patient corresponds to a row in the `description.csv` file, as detailed in Table 1. For each patient, we collected the *PSA* and *Age* information at the time of the initial benign diagnosis, which we do not share for anonymization reasons. The *Positive* column indicates the presence of PCa within the 30 months of follow-up after benign diagnosis, denoted by a value of *1* (one) or *0* (zero) if subsequent diagnoses remain benign over at least the following eight years. In cases of positive diagnoses, the *ISUP* column reflects the ISUP grade of the late detected cancer, ranging from *0* (zero) to *1,2,3,4,5* (one, two, three, four, five) depending on the aggressiveness. Each positive patient was matched with up to three negative control patients for comparative analysis.

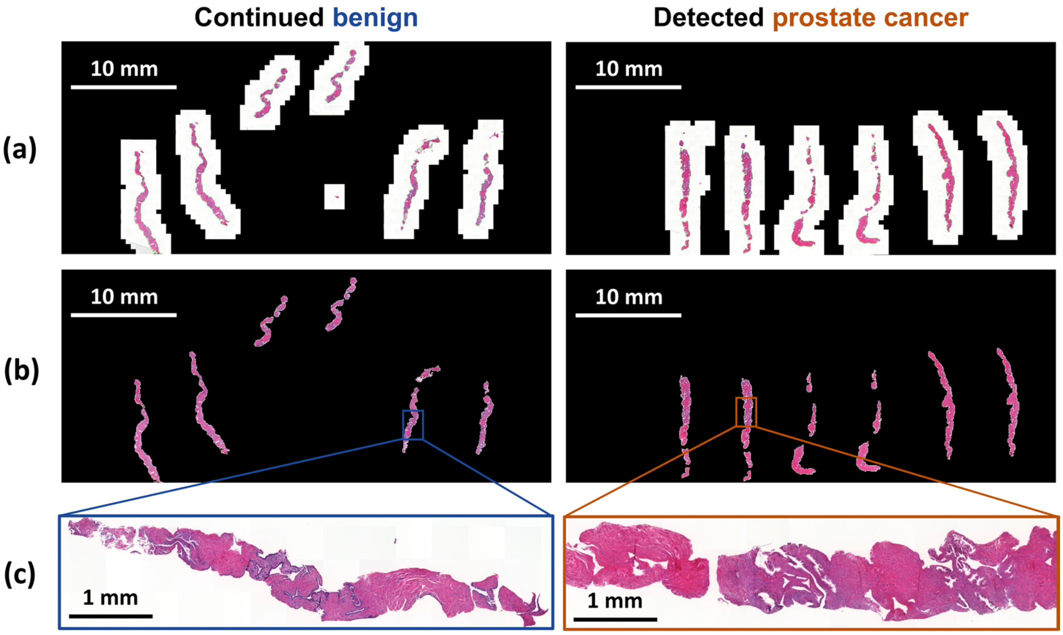

Fig. 2 Example data for one of the slides from a patient that continued being benign or had late detection of prostate cancer. **(a)** Original slide and **(b)** stitched detected tissue[17]. **(c)** Zoom in of a region.

The data is packaged into a `undetected_prostate_cancer.zip` ZIP file. The ZIP file is split into 50 GB parts for convenience.

Technical Validation

The dataset was acquired during routine clinical practice. Upon initial diagnosis, all collected patients underwent evaluation by a histopathologist, establishing a benign diagnosis. Subsequent evaluations by histopathologists upon re-biopsies were conducted to determine the absence or presence of cancer and, in the latter case, to assign a grade. Therefore, the dataset establishes a link between between initial benign diagnoses and potential subsequent cancer outcomes. The multiple follow-up evaluations contribute to the accuracy and reliability of the data.

The clinical variables were collected in a paired manner, ensuring that patients, regardless of their eventual cancer status, are not anticipated to exhibit significantly different years of diagnosis, age, or PSA levels at the time of the first negative biopsy. As depicted in Fig. 1c, which illustrates the distribution of age and PSA levels for both negative and positive cases, the clinical variables alone are insufficient to distinguish between the two groups. These observations were confirmed by the multivariate logistic regression, yielding $R^2 = 0.022$ and p-values for both independent variables well above the conventional significance level of 0.05. Finally, the $AUC = 0.59$ was only slightly above random chance. This highlights the necessity for more advanced computational methods to identify potential changes present at the initial diagnosis that could be indicative of later cancer diagnosis.

To assess the suitability of the dataset with deep learning workflows, we conducted initial preprocessing steps for tissue detection and segmentation. Figure 2b presents the detected tissue for two examples of slides, one from a patient whose condition remained benign and another from a patient later diagnosed with PCa. This analysis facilitated the identification and elimination of slides with corrupted files or undetected tissue. Notably, no further patients were excluded, only specific samples were excluded based on the preprocessing outcome.

Usage Notes

MRXS files can be conveniently inspected using compatible viewer software, such as QuPath[19], ImageJ[20] or TissUUmaps 3[21]. We have provided example data from TissUUmaps in Fig. 2 and an interactive view accessible at https://undetected-prostate-cancer.serve.scilifelab.se/.

The data is readily available to be used in deep learning frameworks. The software utilized for WSI tissue detection and segmentation[17] also includes tiling and feature extraction modules that facilitate image classification. Downsampled versions of the images enable access to lower levels of the resolution pyramid depending on the application. The dataset was previously explored in a weakly-supervised learning framework[16].

Code availability

MRXS files were anonymized using the module in https://github.com/bgilbert/anonymize-slide. The WSI segmentation module proposed in CLAM[17] was used for preprocessing the slides https://github.com/mahmoodlab/CLAM. The images were inspected using TissUUmaps https://tissuumaps.github.io/.

Received: 20 June 2024; Accepted: 5 March 2025;
Published online: 11 March 2025

References

1. Sung, H. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**, 209–249, https://doi.org/10.3322/caac.21660 (2021).
2. Andreoiu, M. & Cheng, L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. *Human pathology* **41**, 781–793, https://doi.org/10.1016/j.humpath.2010.02.011 (2010).
3. Thompson, I. M. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤ 4.0 ng per milliliter. *New England Journal of Medicine* **350**, 2239–2246, https://doi.org/10.1056/nejmoa031918 (2004).
4. Schröder, F. H. Screening and prostate-cancer mortality in a randomized European study. *New England journal of medicine* **360**, 1320–1328, https://doi.org/10.1056/nejmoa0810084 (2009).
5. Norris, J. M. What type of prostate cancer is systematically overlooked by multiparametric magnetic resonance imaging? An analysis from the PROMIS cohort. *European urology* **78**, 163–170, https://doi.org/10.1016/j.eururo.2020.04.029 (2020).
6. Grefve, J. Histopathology-validated gross tumor volume delineations of intraprostatic lesions using PSMA-positron emission tomography/multiparametric magnetic resonance imaging. *Physics and Imaging in Radiation Oncology* **31**, 100633, https://doi.org/10.1016/j.phro.2024.100633 (2024).
7. Heidenreich, A. EAU guidelines on prostate cancer. *European urology* **53**, 68–80, https://doi.org/10.1016/j.eururo.2007.09.002 (2008).
8. Van Leenders, GeertJ. L. H. The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma. *The American journal of surgical pathology* **44**, e87, https://doi.org/10.1097/pas.0000000000001497 (2020).
9. Tosoian, J. J. Active surveillance of grade group 1 prostate cancer: long-term outcomes from a large prospective cohort. *European urology* **77**, 675–682, https://doi.org/10.1016/j.eururo.2019.12.017 (2020).
10. Bjurlin, M. A. Optimization of initial prostate biopsy in clinical practice: sampling, labeling and specimen processing. *The Journal of urology* **189**, 2039–2046, https://doi.org/10.1016/j.juro.2013.02.072 (2013).
11. Sandgren, K. Histopathology-validated lesion detection rates of clinically significant prostate cancer with mpMRI, [68Ga] PSMA-11-PET and [11C] Acetate-PET. *Nuclear medicine communications* **44**, 997–1004, https://doi.org/10.1097/mnm.0000000000001743 (2023).
12. Halin, S., Hammarsten, P., Adamo, H., Wikström, P. & Bergh, A. Tumor indicating normal tissue could be a new source of diagnostic and prognostic markers for prostate cancer. *Expert opinion on medical diagnostics* **5**, 37–47, https://doi.org/10.1517/17530059.2011.540009 (2011).
13. Bulten, W. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine* **28**, 154–163, https://doi.org/10.1038/s41591-021-01620-2 (2022).
14. Liu, B. *et al*. Using deep learning to detect patients at risk for prostate cancer despite benign biopsies, *Iscience***25**, https://doi.org/10.1016/j.isci.2022.104663 (2022).
15. Partin, A. W. Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. *The Journal of urology* **192**, 1081–1087, https://doi.org/10.1016/j.juro.2014.04.013 (2014).
16. Chelebian, E. *et al*. Discovery of tumour indicating morphological changes in benign prostate through AI, medRxiv, https://doi.org/10.1101/2024.06.18.24309064 (2024).
17. Lu, M. Y. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**, 555–570, https://doi.org/10.1038/s41551-020-00682-w (2021).
18. Chelebian, E., Wählby, C. & Bergh, A., Undetected cancer in clinical prostate biopsies, Swedish National Data Service (SND), https://doi.org/10.57804/epa0-8v59 (2024).
19. Bankhead, P. *et al*. QuPath: Open source software for digital pathology image analysis, *Scientific reports* **7**, https://doi.org/10.1038/s41598-017-17204-5 (2017).
20. Schindelin, J. *et al*. Fiji: an open-source platform for biological-image analysis, *Nature methods* **9**, 676–682, https://doi.org/10.1038/nmeth.2019 (2012).
21. Pielawski, N. *et al*. TissUUmaps 3: Improvements in interactive visualization, exploration, and quality assessment of large-scale spatial omics data. *Heliyon* **9**, https://doi.org/10.1016/j.heliyon.2023.e15306 (2023).

Acknowledgements

This research was funded by the European Research Council via ERC Consolidator grant CoG 682810 and Technology Development project from SciLifeLab to C.W and grant no 21-1856 from the Swedish Cancer Society to A.B. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre. Open access funding is provided by Uppsala University.

Author contributions

Conceptualization: E.C., C.W. and A.B.; Data curation: E.C., P.A. and A.B.; Formal analysis: E.C.; Funding acquisition: C.W. and A.B.; Investigation: E.C. and A.B.; Methodology: E.C. and A.B.; Project administration: C.A., C.W. and A.B.; Resources: P.A., H.J. and A.B.; Software: E.C. and C.A.; Supervision: C.A., C.W. and A.B.; Validation: E.C. and C.A.; Visualization: E.C., C.A. and C.W.; Writing – original draft: E.C and A.B.; Writing – review & editing: C.A., H.J., P.A. and C.W.

Funding

Open access funding provided by Uppsala University.

Competing interests

C.W. is on the advisory board of Navinci Diagnostics, Sweden. The rest of the authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.C.

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