nature medicine

Article

https://doi.org/10.1038/s41591-025-03735-2

Early detection of multiple cancer types using multidimensional cell-free DNA fragmentomics

Received: 23 August 2024

Accepted: 24 April 2025

Published online: 27 May 2025

Check for updates

Hua Bao[1,21], Shanshan Yang[1,21], Xiaoxi Chen[1,21], Guangqiang Dong[2], Yuan Mao[3], Shuyu Wu[1], Xi Cheng[1], Xuxiaochen Wu[1], Wanxiangfu Tang[1], Min Wu[1], Shiting Tang[1], Wenhua Liang[4], Zheng Wang[5], Liu Yang[6], Jiaqi Liu[7], Tao Wang[8], Bingzhong Zhang[9], Kuirong Jiang[10], Qin Xu[11], Jierong Chen[12], Hairong Huang[13], Junjie Peng[14], Xiaomeng Xia[15], Yumei Wu[16], Shun Xu[17], Ji Tao[18], Li Chong[19], Dongqin Zhu[1], Ruowei Yang[1], Shuang Chang[1], Peng He[1], Xiuxiu Xu[1], JinPeng Zhang[1], Yi Shen[1], Ya Jiang[1], Sisi Liu[1], Xian Zhang[1], Xue Wu[1], Xiaonan Wang[1] & Yang Shao[1,20] ✉

The multicancer early detection (MCED) test has the potential to enhance current cancer-screening methods. We evaluated a new MCED test that analyzes plasma cell-free DNA using genetic- and fragmentomics-based features from whole-genome sequencing. The present study included an internal validation cohort of 3,021 patients with cancer and 3,370 noncancer controls, and an independent cohort of 677 patients with cancer and 687 noncancer individuals. The results demonstrated an overall sensitivity of 87.4%, specificity of 97.8% and tissue-of-origin prediction accuracy of 82.4% in the independent validation cohort. Preliminary results from a prospective study of 3,724 asymptomatic participants showed a sensitivity of 53.5% (predominantly early stage cancers) and specificity of 98.1%. These findings indicate that the MCED test has strong potential to improve early cancer detection and support clinical decision-making.

Currently, noncommunicable diseases are the leading causes of global mortality, accounting for over 36 million deaths annually[1]. Cancer deaths ranked second only to heart disease, causing ~10 million mortalities with ~19.3 million diagnoses in 2020 (refs. 2,3). Most cancer incidents and fatalities are concentrated within ten common cancer types, which are responsible for ~60% of incidences and ~70% of mortalities[3].

Early detection of cancer is essential for effective cancer treatment and can lead to a more favorable prognosis for patients[4]. However, a limited number of early screening options exist and target only a specific type of cancer[4]. For example, the United States Preventive Services Task Force (USPSTF) currently recommends four cancer early screening techniques including mammography tests for breast cancer[5], high-risk human papillomavirus tests for cervical cancer[6], low-dose computed tomography tests for lung cancer[7] and stool-based tests and visualization tests for colorectal cancer[8]. Despite their potential benefits, the implementation of early detection tests in clinical practice, particularly for multicancer screening, has been hindered by several factors, including the time required to perform screening, scheduling challenges, concern over test invasiveness and pain, radiation exposure, fear of the test, discomfort and financial costs[9–12]. Hence, there is an urgent need for a noninvasive, cost-effective, multicancer early detection (MCED) test that detects signals for multiple cancers from cell-free DNA (cfDNA) or other circulating analytes in the blood shed by tumors with a high sensitivity to advance the field of oncology.

cfDNA, which is fragmented DNA released into the circulation by cancer cells, carries a biosignature for the tissue of origin and has become a promising noninvasive liquid-biopsy biomarker for early

A full list of affiliations appears at the end of the paper. ✉e-mail: yang.shao@geneseeq.com

https://doi.org/10.1038/s41591-025-03735-2

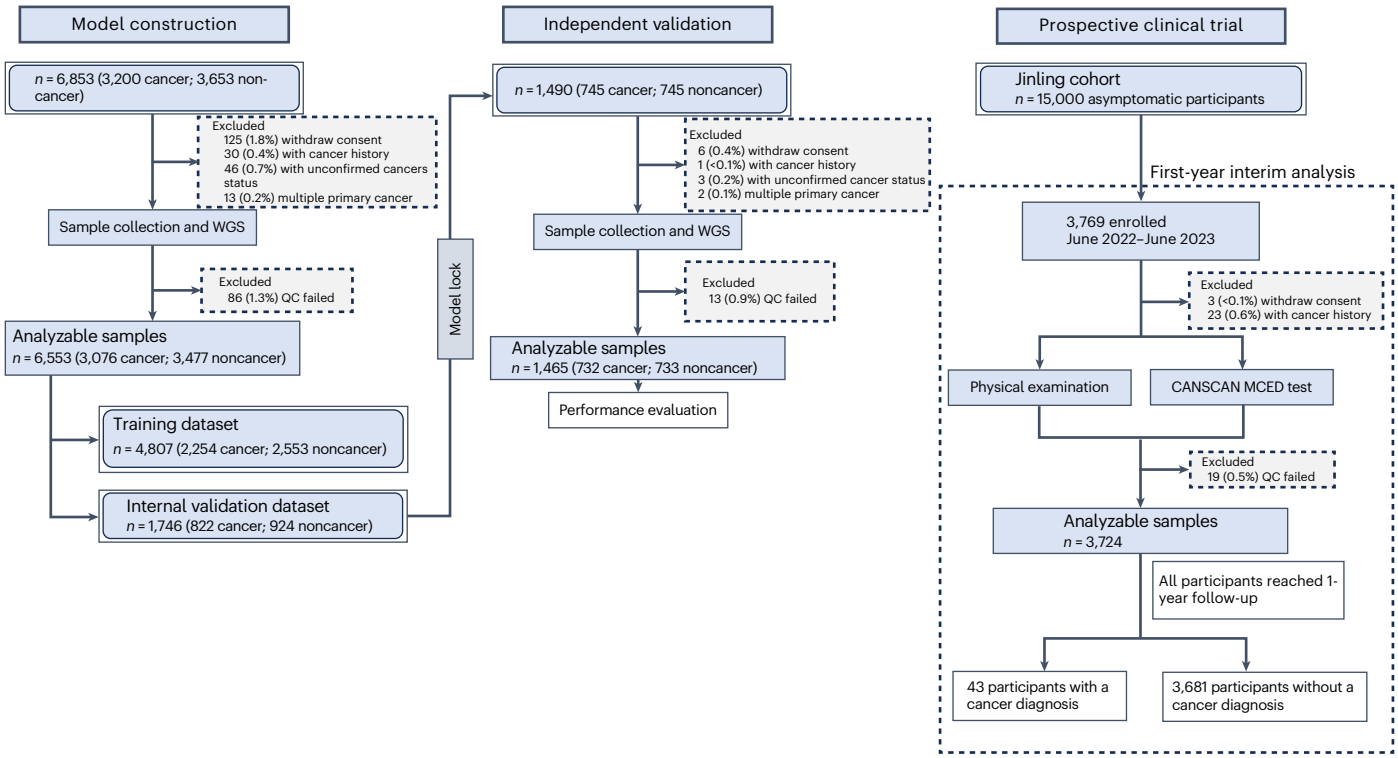


Fig. 1 | Flow chart of study participants. Left: model construction including 6,853 participants, with 6,553 remaining after exclusions. The remaining participants were divided into training (n = 4,807) and internal validation (n = 1,746) datasets. Middle: independent validation used 1,465 of 1,490 participants after applying the same exclusion criteria. Right: the prospective cohort study, Jinling cohort, aiming to enroll 15,000 asymptomatic participants, with a first-year interim analysis of 3,724 participants. QC, quality control.

cancer detection[13–15]. Although traditional tumor somatic mutation calling-based methods show poor predictive power as a result of mutation variability and low tumor concentration, more recent DNA methylation and circulating (ct)DNA fragmentomic signatures have shown great potential for early cancer detection[16–23]. Some MCED tests have also been started or are in preparation for population-based MCED blood test screening[24–26].

Previous studies demonstrated high sensitivities for tests focused only on six or fewer cancer types, whereas studies encompassing multiple cancer types had lower aggregate sensitivities (<60%)[27]. This discrepancy is further exacerbated in population-based studies of healthy individuals[28]. These findings have raised concerns among distributors who fear that reduced sensitivity could cause a false sense of security, potentially dissuading patients from pursuing further screenings[24].

In the present study, we have developed and validated a noninvasive MCED blood test capable of identifying cancers across 13 cancer types and localizing the tissue of origin. The cancers targeted for detection include cancers of the breast, cervical, colorectal, endometrial, esophageal, gastric, liver, lung, ovarian, pancreatic, prostate, bile duct and lymphoma. Collectively, these cancers represent 66.6% of new cancer incidences and 74.0% of cancer-related mortalities worldwide. Our test harnesses plasma cfDNA and employs characteristics of genome-wide fragmentomics to enhance detection sensitivity through the integration of multidimensional signals. The development process encompassed retrospective model training and validation, followed by prospective independent validation, resulting in a robust multicancer detection blood test model that has undergone clinical validation. Furthermore, we present the interim analysis results from a large-scale, prospective population study. The present study investigates the feasibility of MCED testing in an outpatient setting, targeting adults aged 45–75 years who are asymptomatic for cancer.

Results

Study participants

In this multiphase study involving multiple Chinese cohorts, participant recruitment and sample collection were conducted systematically (Fig. 1). During the model construction phase, we obtained qualified blood samples from a meticulously curated cohort of 3,076 patients with cancer, representing 13 distinct cancer types and 3,477 age-matched healthy controls. These samples were stratified into two sets: a training set (n = 4,807) and an internal validation set (n = 1,746). Subsequently, we conducted an independent validation with a prospectively enrolled cohort of 1,465 participants in an age-matching fashion, comprising 732 patients with cancer and 733 noncancer individuals. The baseline demographics were comparable across the training, internal validation and independent validation cohorts, as detailed in Table 1.

The third phase, currently in progress, offers an interim analysis from an ongoing prospective cohort study, the Jinling study (NCT06011694). This study is actively enrolling participants between the ages of 45 and 75 years who are asymptomatic for cancer, initiated in June 2022. As of June 2023, a total of 3,769 participants had been enrolled; 26 were subsequently excluded as a result of either withdrawal from the study or the presence of a pre-existing cancer diagnosis that was undisclosed at the time of enrollment and 19 were excluded because of the unavailability of evaluable samples. Consequently, MCED test results and physical examination reports were successfully obtained and analyzed for the remaining 3,724 participants.

Development and validation of the blood-based MCED test

Cancer signal detection. The MCED test integrates a sophisticated framework that analyzes a broad set of genomic and fragmentomics features from cfDNA, derived from low-coverage whole-genome sequencing (WGS). The model is structured around two core classifiers: the detection-of-cancer (DOC) classifier, tasked with confirming the

https://doi.org/10.1038/s41591-025-03735-2

Table 1 | Demographic and clinical characteristics of model construction and validation participants

	Training		Internal validation		Independent validation	
	Cancer	Noncancer	Cancer	Noncancer	Cancer	Noncancer
	n=2,254	*n*=2,553	*n*=822	*n*=924	*n*=732	*n*=733
Age (years), median (IQR)	59 (51–67)	56 (52–62)	59 (52–68)	56 (52–62)	59 (51–67)	57 (52–62)
Sex, *n* (%)						
Male	1,112 (49.3)	1,121 (43.9)	429 (52.2)	398 (43.1)	350 (47.8)	326 (44.5)
Female	1,142 (50.7)	1,432 (56.1)	393 (47.8)	526 (56.9)	382 (52.2)	407 (55.5)
TNM stage, *n* (%)						
I	876 (38.9)	–	274 (33.3)	–	227 (31.0)	–
II	552 (24.5)	–	222 (27.0)	–	222 (30.3)	–
III	504 (22.4)	–	200 (24.3)	–	184 (25.1)	–
IV	207 (9.2)	–	81 (9.9)	–	70 (9.6)	–
Uncertain	115 (5.1)	–	45 (5.5)	–	29 (4.0)	–
Cancer type, *n* (%)						
Bile duct	55 (2.4)	–	19 (2.3)	–	20 (2.7)	–
Breast	144 (6.4)	–	53 (6.4)	–	47 (6.4)	–
Cervical	63 (2.8)	–	21 (2.6)	–	19 (2.6)	–
Colorectal	260 (11.5)	–	111 (13.5)	–	96 (13.1)	–
Endometrial	180 (8.0)	–	59 (7.2)	–	54 (7.4)	–
Esophageal	144 (6.4)	–	50 (6.1)	–	44 (6.0)	–
Gastric	139 (6.2)	–	46 (5.6)	–	39 (5.3)	–
Liver	231 (10.2)	–	78 (9.5)	–	74 (10.1)	–
Lung	603 (26.8)	–	232 (28.2)	–	200 (27.3)	–
Lymphoma	125 (5.5)	–	42 (5.1)	–	38 (5.2)	–
Ovarian	72 (3.2)	–	25 (3.0)	–	21 (2.9)	–
Pancreatic	126 (5.6)	–	42 (5.1)	–	39 (5.3)	–
Prostate	112 (5.0)	–	44 (5.4)	–	41 (5.6)	–

Note: TNM classification refers to the tumor, node, metastasis staging of cancer severity.

presence of cancer, and the tissue-of-origin (TOO) classifier, responsible for pinpointing the primary site of the malignancy. We initially trained the DOC classifier on a specific training dataset, with subsequent refinements to meet two distinct specificity levels, as corroborated by our internal validation. The standard specificity threshold (SST) was established at 98.9% (95% confidence interval (CI) = 98.0–99.4%) and achieved a sensitivity of 86.5% (95% CI = 84.0–88.7%; Fig. 2a). Under the high specificity threshold (HST), the sensitivity was 82.4% (95% CI = 79.6–84.8%; Supplementary Table 1) and the specificity was high at 99.9% (95% CI = 99.4–100.0%).

The performance was further corroborated through an independent validation study. Within this independent dataset, the MCED test maintained a commendable sensitivity of 87.4% (95% CI = 84.8–89.6%) and specificity of 97.8% (95% CI = 96.5–98.7%) at the SST. When applying the HST, the test's sensitivity was noted at 82.5% (95% CI = 79.6–85.1%), with the specificity reaching 98.9% (95% CI = 97.9–99.4%). These results from the independent validation reinforce the MCED test's capability to consistently identify the presence of cancer with high accuracy.

The MCED test demonstrates a high sensitivity for detecting early stage cancers, with a sensitivity of 79.3% (95% CI = 73.6–84.1%) and 86.9% (95% CI = 81.9–90.7%) for stages I and II, respectively, in the independent validation set at the SST (Fig. 2b). As expected, sensitivity of cancer signal detection increased with increasing stage, with 92.4% (95% CI = 87.6–95.4%) and 97.1% (95% CI = 90.2–99.2%) for patients in stages III and IV at SST, respectively. Notably, this sensitivity remains high at 71.8% (95% CI = 65.6–77.3%) and 83.3% (95% CI = 77.9–87.7%) for patients in stages I and II when applying the HST, indicating the test's strong performance in accurately identifying stage I cancers under more stringent conditions (Supplementary Table 2). The overall sensitivity at SST for patients in either stages I–II or stages I–III was 82.7% (95% CI = 79.1–85.7%) and 85.2% (95% C = 82.4–87.6%), respectively.

The diagnostic test under investigation demonstrates a notable ability to identify signals across a spectrum of cancer types (Fig. 2c) and the comprehensive sensitivity data, stratified by cancer type and clinical stage, are provided in Supplementary Table 3. Notably, the sensitivity of the test for detecting liver and bile duct cancers is 100.0%, suggesting an exceptionally high level of accuracy for these malignancies. In addition, the test demonstrates robust performance in the identification of lung and colorectal cancers, with sensitivities of 94.5% and 82.3%, respectively. The test's effectiveness extends to several cancers that currently lack established screening protocols. Pancreatic cancer, characterized by its typically asymptomatic early stages and rapid progression, demonstrated a sensitivity of 76.9% in the detection of 39 evaluated cases, which breaks down into 58.3% for stage I, 86.7% for stage II and 77.8% for stage III. Furthermore, ovarian cancer detection via the test is highly effective, with a sensitivity of 90.5%.

Prediction of tissue origin. The efficacy of a blood-based multicancer detection test is greatly enhanced by its ability to accurately predict the TOO, which is essential for guiding subsequent diagnostic procedures. The study establishes the accuracy of TOO prediction for ten cancer

https://doi.org/10.1038/s41591-025-03735-2

a

	n	Cases	Detection of cancer	
			Sensitivity (95% CI)	Specificity (95% CI)
Internal validation	1,746	822	86.5% (84.0–88.7%)	98.9% (98.0–99.4%)
Independent validation	1,465	732	87.4% (84.8–89.6%)	97.8% (96.5–98.7%)

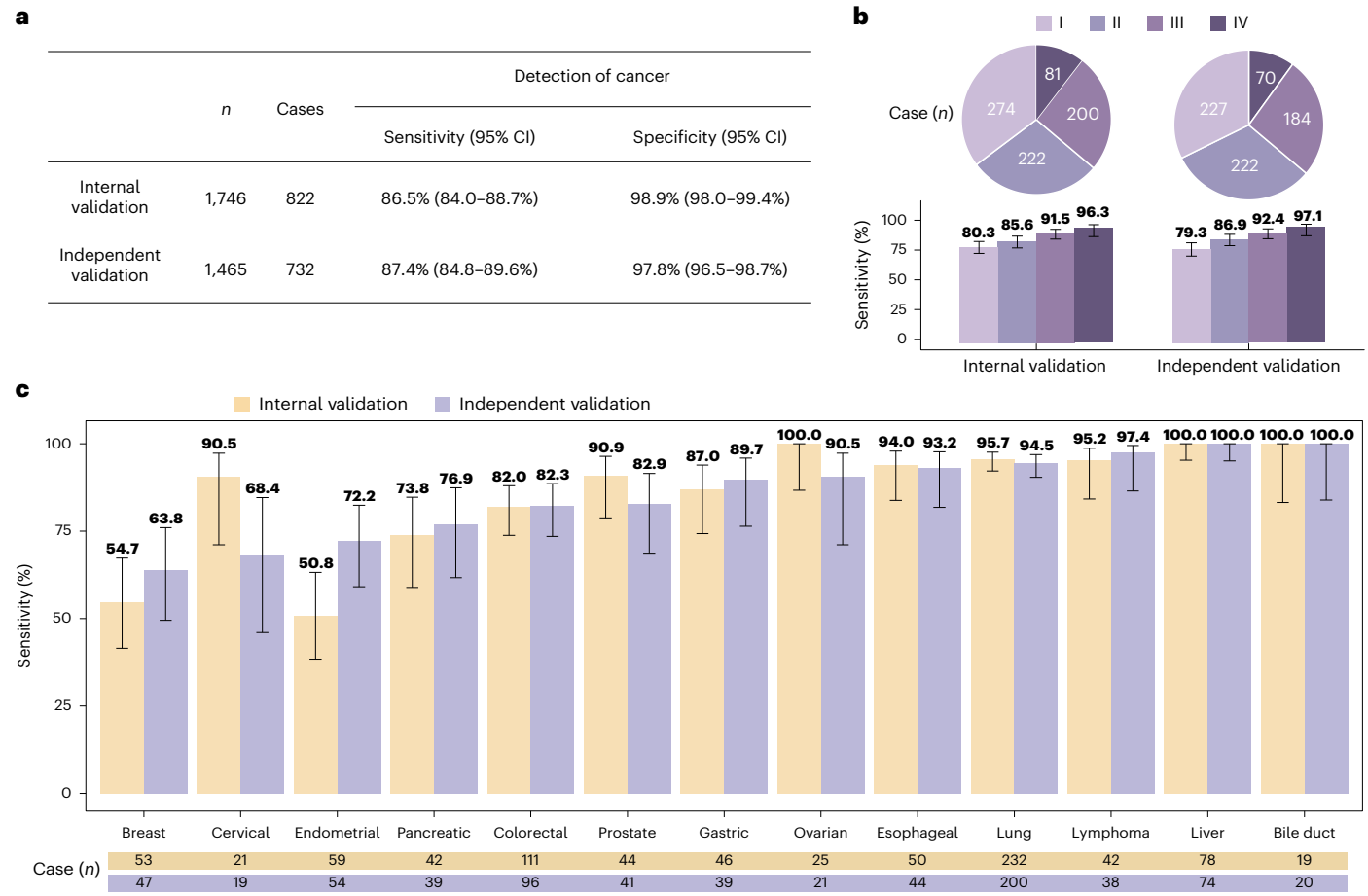


Fig. 2 | Performance of the MCED test in detecting cancer signals. a, Sensitivity and specificity of the MCED test in both internal and independent validation sets at two different specificity thresholds. **b**, Distribution of cancer stages within the internal validation cohort ($n = 777$) and the independent validation cohort ($n = 703$), presented as frequencies. Bar graphs depict the sensitivity of the MCED test across different cancer stages for each validation cohort, with error bars representing the 95% CIs of sensitivity. Statistical analysis was performed on independent patient samples within each stage group. **c**, Bar graphs displaying a comparative analysis of the MCED test sensitivity across various cancer types. Data are shown for both the internal validation cohort and the independent validation cohort, with cases for each presented distinctly as indicated in the color-coded legend. Error bars on the bars represent the 95% CIs of sensitivity. Sensitivity calculations for each cancer type were based on data from individual patient samples contributing to the case numbers denoted in the legend.

types. For the primary tissue origin prediction (TOP1), accuracies of 82.3% (95% CI = 79.2–85.0%) and 83.5% (95% CI = 80.6–86.4%) were reported for internal and independent validation sets, respectively (Fig. 3a,b). When considering the top two tissue origin prediction (TOP2), the accuracy increased to 90.7% (95% CI = 88.6–92.9%) for the internal set and 91.7% (95% CI = 89.5–93.9%) for the independent set (Extended Data Fig. 1). The TOO prediction accuracy showed further improvement in the independent validation set at HST (Supplementary Table 1), with TOP1 accuracy reaching 83.9% (95% CI = 81.0–86.8%) and TOP2 accuracy achieving 91.9% (95% CI = 89.5–94.0%). The classifier demonstrated robust performance across the various cancer stages included in the study (Supplementary Tables 4 and 5).

The classifier's performance in identifying the TOP1 across various cancer types demonstrates notable variability, which may influence the strategic approach to subsequent diagnostic and therapeutic interventions. The TOP1 accuracy rates were high for colorectal cancer, lung cancer, liver cancer and lymphoma cancer in both validation cohorts (Fig. 3c,d). These cancers showed robust detection capabilities, suggesting a higher reliability of the test in these contexts. Conversely, pancreatic and gastric cancers exhibited the lowest TOP1 accuracy rates at 50.0% and 45.7% in independent validation, respectively, indicating potential challenges in the test's ability to correctly identify the tissue of origin in these cases.

MCED test in asymptomatic cohort screening. The Jinling study is a prospective, multicenter, observational cohort designed to enroll 15,000 individuals aged 45–75 years at two physical examination sites. Participants undergo annual routine physical examinations and MCED testing for 3 years, followed by a 2-year electronic health record surveillance period. Participants were excluded if they had a history of cancer or current diagnosis of cancer, or had received an organ transplant or prior nonautologous (allogeneic) bone marrow or stem-cell transplant. Participants provided written informed consent.

Of the 3,724 participants of the Jinling cohort who were available for analysis, the median age was 56 years, with an interquartile range (IQR) of 52–61 years. Of these participants, 2,435 (65.4%) were female and 1,289 (34.6%) male. A total of 762 participants (20.5%) reported being current or former smokers (Supplementary Table 6). In this interim analysis, all patients were included with at least a 1-year follow-up period (Extended Data Fig. 2). Overall, a total of 43 cases of cancer were confirmed, resulting in a sensitivity of 53.5% (Fig. 4a) at SST and 37.2% at HST (Supplementary Table 7). A predominant majority of these cases were diagnosed at early stages—specifically, high-risk pre-cancer lesions, stage 0, I or II—accounting for 93.0% (40 of 43) of the instances. Among the 39 cases with malignant cancer excluding pre-cancerous lesions, the MCED test successfully identified cancer in 21 individuals, yielding a sensitivity of 53.8% (Fig. 4a) at SST and 35.9%

https://doi.org/10.1038/s41591-025-03735-2

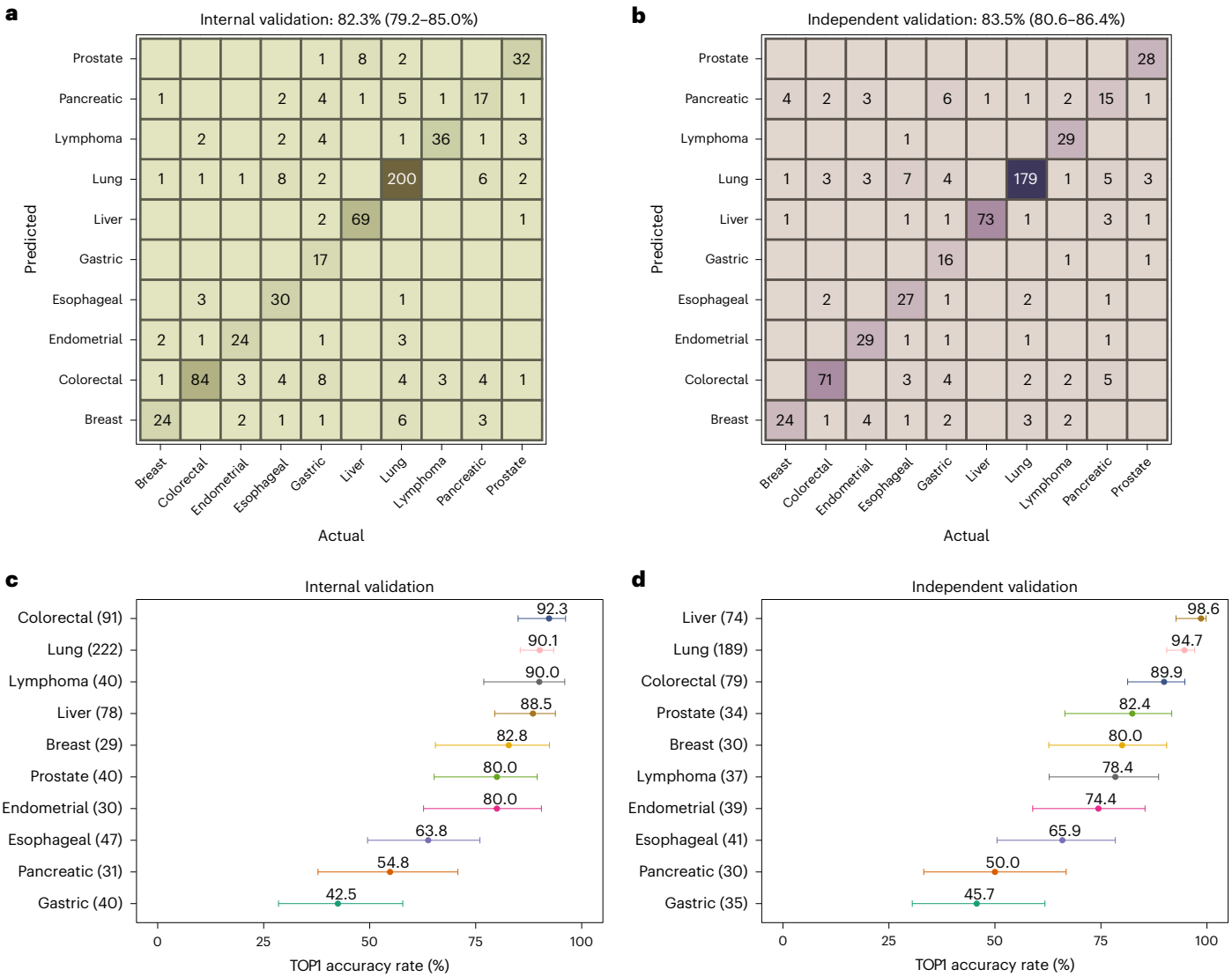


Fig. 3 | TOP1 accuracy. a, Confusion matrix displaying the predicted and actual cancer types for the internal validation set. **b**, Confusion matrix displaying the predicted and actual cancer types for the independent validation set. **c**, TOP1 accuracy rates for different cancer types in the internal validation set, with error bars representing 95% CIs of accuracy. The number of patients in each cancer type is indicated in parentheses. **d**, TOP1 accuracy rates for different cancer types in the independent validation set, with error bars representing 95% CIs of accuracy. The number of patients in each cancer type is indicated in parentheses. This analysis is based on independent patient samples, with the unit of study being the individual patient. No technical or biological replicates were performed at the individual patient sample level for this analysis. The *n* values represent counts of unique patients.

at HST (Supplementary Table 7). For the 13 designated cancer types, the sensitivity reached 62.1% (18 of 29). The specificity of the test was reported at 98.1% (95% CI = 97.6–98.5%). The sensitivity varied according to the stage of cancer and the primary cancer site (Fig. 4b,c), with stage-specific sensitivities of 44.8% (13 of 29) for stage 0 or I and 85.7% (6 of 7) for stage II. Overall, the MCED test identified a cancer signal in 92 instances, with a positive predictive value (PPV) of 25.0% (17.3–34.7%) in this setting. The negative predictive value (NPV) was exceedingly high at 99.4% (99.2–99.6%).

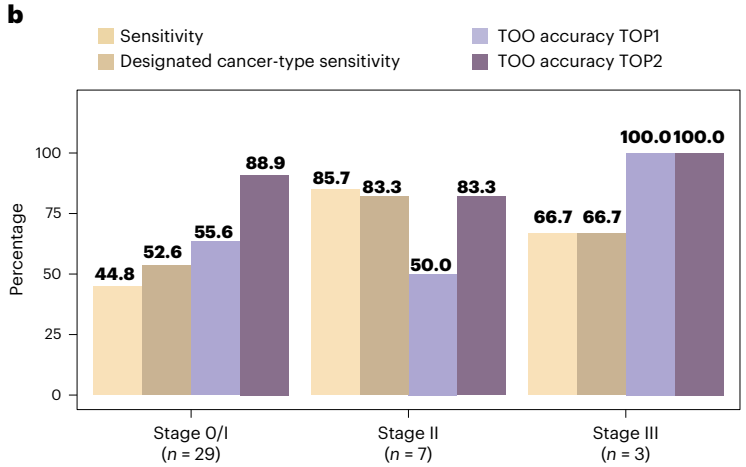
Participant outcomes from physical examinations were recorded, with an established criterion for cancer-related physical examination abnormalities to mitigate the variation in clinical practice across recruiting sites (Extended Data Fig. 3 and Supplementary Table 6). The most common site of cancer-specific abnormalities was the breast (23.6%), followed by thyroid (13.3%) and lung (9.2%). The incidence of cancer diagnoses varied across different clusters of cancer-specific abnormalities (Supplementary Table 8). Specifically, cancer was identified in 1 of 28 participants (3.6%) within the liver abnormality cluster, 11 of 340 participants (3.2%) in the lung abnormality cluster, 2 of 92 participants (2.2%) in the colorectal abnormality cluster, 11 of 574 participants (1.9%) in the breast abnormality cluster and 3 of 110 participants (2.7%) in the cervical abnormality cluster.

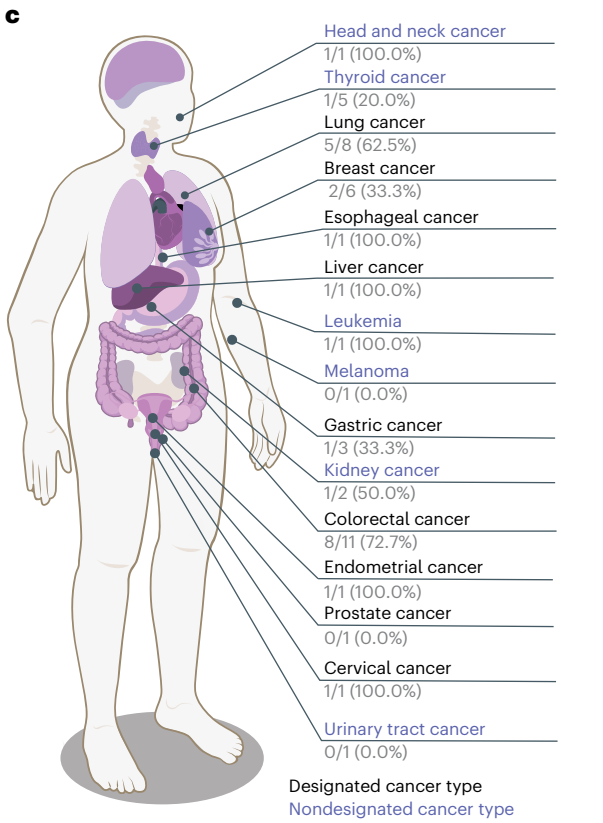
In examining the post-test cancer diagnostic probability—representing the post-test probability that cancer is present after a positive MCED test—of cancer across various abnormality clusters, a distinct pattern emerged (Fig. 4d). For lung, colorectal, breast, cervical and liver cancers, the post-test diagnostic probability increased notably. On the other hand, the post-test cancer exclusion assurance, which indicates the post-test probability of cancer being present after a negative test result, remained consistently low for these clusters. In the cluster of breast cancer screening, the pre-test probability was established at 1.9%, whereas the diagnostic probability after a positive test outcome escalated to 33.3%. This elevation signals a considerable increase in the likelihood of cancer in the event of a positive test. Conversely, the post-test cancer exclusion assurance was reported at 1.1%, indicating a minimal chance of cancer present after a negative result. For the

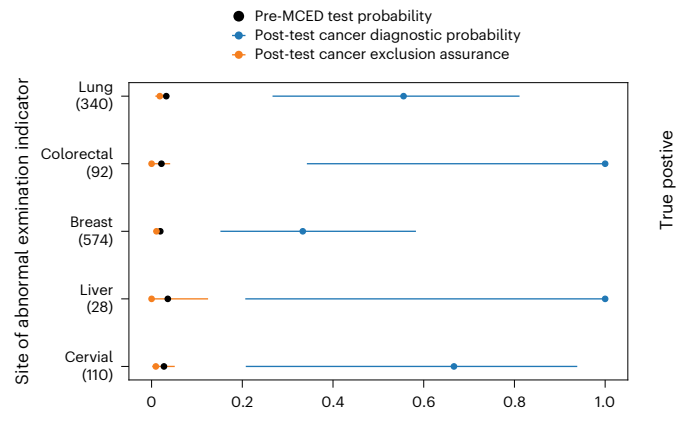
https://doi.org/10.1038/s41591-025-03735-2

a

Jinling cohort	
Sample size	3,724
Total cases	43
Metric (*n*/*N*; %, 95% CI)	
Sensitivity	23/43; 53.5 (38.9–67.5%)
Sensitivity excluding pre-cancer[a]	21/39; 53.8 (38.6–68.4%)
Designated cancer-type sensitivity[b]	18/29; 62.1 (44.0–77.3%)
Specificity	3,612/3,681; 98.1 (97.6–98.5%)
PPV	23/92; 25.0 (17.3–34.7%)
NPV	3,612/3,632; 99.4 (99.2–99.6%)
TOO accuracy TOP1	12/19; 63.2 (42.1–84.2%)
TOO accuracy TOP2	17/19; 89.5 (73.7–100.0%)







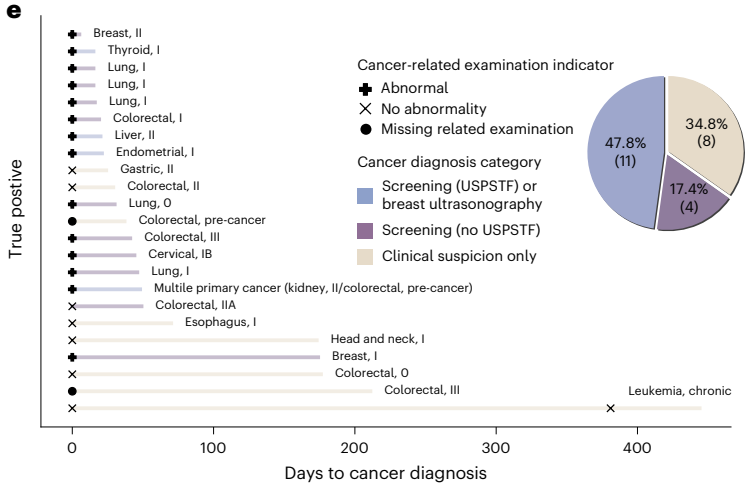


Fig. 4 | Performance of the MCED test in the Jinling cohort. a, Performance metrics for the MCED test, including sensitivity for all cancer types and the 13 designated cancer types, specificity, PPV, NPV and TOP1 and TOP2 accuracy. [a]Sensitivity without pre-cancer is the sensitivity of the test specifically for detecting confirmed malignant neoplasms (cancers), calculated by excluding pre-cancerous lesions from the numerator (true positives) and the denominator (total cancer cases). [b]Designated cancer-type sensitivity is the sensitivity for the confirmed malignant neoplasms in specific cancer types for which the test was designed. **b**, Sensitivity of the MCED test by cancer stage, including stages 0–I, II and III cancers, as well as TOP1 and TOP2 accuracy for each stage. **c**, Sites of all 43 cancer instances in the Jinling cohort, indicating the number of cases and the percentage of total cases detected by the MCED test for each cancer type. **d**, Pre- and post-test probabilities of cancer by the site of abnormal physical examination indicators, including pre-MCED test probability, post-test cancer diagnostic probability and post-test cancer exclusion assurance for lung, colorectal, breast, liver and cervical cancers. **e**, Time to cancer diagnosis and cancer-related examination indicators for true positive MCED test results, showing days to cancer diagnosis for each true positive case, whether the cancer was associated with an abnormal physical examination, no abnormality or missing related examination, and the screening category for each true positive case.

cervical cancer screening subgroup, the preliminary probability of cancer was slightly higher at 2.7% and the post-test diagnostic probability was 66.6%, while maintaining a cancer exclusion assurance of 0.9%. Colorectal and liver cancer clusters showed a pre-test probability of 2.2% and 3.6%, respectively, with both reaching a post-test cancer diagnostic probability of 100%, indicating that all positive tests corresponded to a cancer diagnosis within the study's context. The post-test cancer exclusion assurance for both was 0.0%, suggesting that a negative test rules out the presence of cancer. For lung cancer, the pre-test probability was 3.2% with a post-test cancer diagnostic probability of 55.6% and the cancer exclusion assurance was 0.8%, aligning with the high probability of ruling out cancer seen in other clusters.

The cancer diagnosis pathways of the Jinling study incorporated physical exams provided by our study, whether or not with a USPSTF recommendation and clinical suspicion through routine clinical care (Fig. 4e and Extended Data Fig. 4). A total of 34.8% (8 of 23) of patients who received a positive MCED result had their cancers undetected through physical examination and 17.4% (4 of 23) had cancers for which there is currently no recommended screening (Fig. 4e and Supplementary Table 9). In addition, the MCED test showed a concordance rate of 61.1% (11 of 18), with cancers identified by breast ultrasonography or USPSTF-recommended screening methods, while identifying 50.0% (4 of 8) of the cases of cancer that were detected by nonstandard screening protocols (Extended Data Fig. 5). Overall, the MCED test succeeded in identifying 48.0% (12 of 25) of the cases of cancer that had not been detected by breast ultrasonography or screening methods recommended by the USPSTF. These findings underscore the potential of the MCED test to improve cancer detection rates, particularly for cancer types that either lack specific screening guidelines or are not detectable using current screening methodologies.

Discussion

Simultaneously screening for multiple cancer types has the potential to revolutionize current screening practices[27]. The specificity and sensitivity metrics of the MCED test are paramount in evaluating its clinical utility and impact. High specificity is critical to minimizing false-positive results, which can lead to unnecessary anxiety, further invasive diagnostic procedures and potentially inappropriate treatment for patients. In the case–control study and interim results of the screening of an asymptomatic cohort, the low false-positive rate ensures that most individuals identified as cancer free truly do not have a malignancy, thereby preserving patient confidence and resource allocation within healthcare systems.

The present study is one of the first large-scale prospective evaluations of an MCED test based on WGS of blood cfDNA. The MCED test utilizes a multidimensional genomic and fragmentomic approach for cfDNA, including the analysis of cfDNA fragmentation patterns, inferred methylation profiles and gene expression indicators. This comprehensive analysis provided by WGS, in contrast to targeted panel methods, enables ongoing updates to the test, allowing for enhancements to its accuracy and efficacy without the need for re-sequencing[22,29–31]. Previous studies, such as the DELFI and GEMINI studies[17,23], also utilized low-coverage WGS of cfDNA and reported high performance, particularly in lung and liver cancers. Similarly, in our independent validation cohort, our DOC classifier achieved high sensitivity (>80%) for these cancers. The false-positive rate (FPR) of MCED has shown notable improvement compared with the 5–15% FPR commonly associated with current standard-of-care screening tests, with many systems achieving an FPR of around ~1%. Despite this advance, the sensitivity of cfDNA tests for early stage cancer detection remains suboptimal. The ability of these tests to identify stage I and II cancers is particularly concerning, with reported sensitivities ranging from 27.5% (95% CI = 25.3–29.8%) to 37.3% (95% CI = 29.8–45.4%) and reaching up to 43.8% (95% CI = 36.8–50.9%) in some studies[16,18,19]. However, the sensitivity of cfDNA assays for multicancer early detection remains suboptimal, particularly for early stage cancers (stages I and II). Sensitivity rates for detecting stage I and II cancers have been reported in the range of 27.5% (95% CI = 25.3–29.8%), 37.3% (95% CI = 29.8–45.4%) to 43.8% (95% CI = 36.8–50.9%)[16,18,19]. This considerable rate of false negatives has raised alarms among private insurers, who worry that such results may impart a false sense of security to patients and potentially deter them from pursuing further necessary screenings[28]. Conversely, the impressive sensitivity of the MCED test in the present study covering 13 cancer types (or 30 American Joint Committee on Cancer (AJCC) cancer types; Supplementary Table 10), particularly at early cancer stages (stages I and II), underscores its capability for early intervention when cancers are most treatable.

Accurate TOO predictions are paramount for guiding subsequent diagnostic and therapeutic pathways, thereby optimizing patient outcomes. The study's demonstration of high TOO prediction accuracy at TOP1 and TOP2 underscores the robustness of the test and its applicability across a variety of cancer types, which includes those currently lacking formal screening protocols. However, it is slightly lower than what has been reported in previous studies: 88.7% for 50 AJCC cancer types[18] and 89.7% for 6 cancer types[19]. It is important to note that these reported TOO accuracies typically account for patients correctly identified as having cancer by the model. In contrast, the sensitivity in our study is notably higher, indicating a broader detection capability that includes patients with lower cancer signals. Consequently, our inclusion criteria might lead to a somewhat lower TOO prediction accuracy but reflects a more comprehensive detection capability.

One of the most compelling aspects of the present study was the screening of an asymptomatic cohort (Jinling study). The PPV achieved was 25%, which is lower than the 38% reported in the PATHFINDER study[24]. However, drawing direct comparisons requires careful consideration of the fundamental differences in study design and cohort characteristics. The PATHFINDER trial enrolled participants with a higher cancer prevalence and utilized the MCED test result to trigger diagnostic workup, systematically investigating participants with positive test results, thus inherently enriching their cohort for cancer diagnoses within the workup pathway. Conversely, the Jinling study adopted a standardized comprehensive physical examination for all participants as the primary screening modality, independent of MCED test outcomes and within a context of a lower cancer prevalence.

Our study demonstrated high sensitivity, highlighting our classifier's ability to detect cancer cases even in populations with lower disease prevalence. This underscores the capacity of our classifier to effectively detect incident cancer cases under real-world screening conditions facilitated by comprehensive physical examinations.

It is particularly relevant to note that the observed invasive cancer incidence within the Jinling cohort was 1.0% (excluding pre-cancer lesions), higher than an age- and gender-adjusted expected 1-year cancer incidence rate of 0.67%, calculated using national registry data[32]. This residual elevation is probably attributable to the comprehensive nature of physical examinations enabling earlier cancer detection and the large proportion (51.5%) of participants with pre-existing elevated cancer risk profiles based on their lifestyle and family history. In alignment with the observational nature of the Jinling study design, cancer outcomes were ascertained through the application of standard-of-care diagnostic pathways at each participating study site, after comprehensive physical examinations of all participants. It is important to note that all participants within the Jinling cohort received appropriate clinical management and treatment according to established standard-of-care procedures, based on the findings from these standardized physical examinations, ensuring ethical clinical practice and appropriate medical follow-up within the study protocol. Variation in clinical practice across study sites was mitigated by an established criterion of cancer-related physical examination abnormalities that followed standards of care according to the China Guideline for Cancer Screening and Chinese Society of Clinical Oncology guidelines. Notably, the compliance rates for screening tests during physical examinations were higher in the Jinling study (Supplementary Table 6) because participants were encouraged to undergo a comprehensive physical examination, which led to a high coverage of cancer incidence detection through physical examinations when comparing with previous trials or population studies[33–35]. The high overall sensitivity, specificity and post-test probabilities for cancer signal detection, reported across cancer abnormality clusters, indicate that the integration of traditional diagnostic methods with advanced genomic data not only reinforces the reliability of the MCED test but also highlights the complementary role that it can play alongside existing screening paradigms.

However, the findings still have some limitations. The study included only 13 cancer types, although these represent most of the common types of cancer in China. In addition, the study was confined to Chinese patients, which constrains the generalizability of the cancer detection rate as a result of the limited ethnic, racial and socioeconomic diversity of the study cohort, as well as the high rate of baseline adherence to cancer screening. Specifically, esophagogastroduodenoscopy for esophageal and gastric cancer screening was not included in our physical examination protocol, because of the low compliance rates for endoscopy in China[36], which may introduce bias by limiting the detection of relative cancers during physical examinations in the interim analysis of the Jinling cohort. Several cancer types had a limited number of samples, which restricted the construction of the tumor-of-origin model and was not included in the TOO prediction. This limitation of the TOO capacity restricts its use in asymptomatic cohorts; however, this can be overcome by updating the model. Finally, the observational nature of the prospective cohort study indicates that further interventional studies are needed.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-03735-2.

References

1. Banatvala, N., Akselrod, S., Bovet, P. & Mendis, S. in *Noncommunicable Diseases* 234–239 (Routledge, 2023).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30 (2020).
3. 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).
4. *Guide to Cancer Early Diagnosis* (World Health Organization, 2017).
5. Siu, A. L. et al. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **164**, 279–296 (2016).
6. Curry, S. J. et al. Screening for cervical cancer: US Preventive Services Task Force recommendation statement. *JAMA* **320**, 674–686 (2018).
7. Krist, A. H. et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA* **325**, 962–970 (2021).
8. Davidson, K. W. et al. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. *JAMA* **325**, 1965–1977 (2021).
9. Patel, M. et al. Hepatocellular carcinoma: diagnostics and screening. *J. Eval. Clin. Pract.* **18**, 335–342 (2012).
10. The National Lung Screening Trial Research et al. Results of initial low-dose computed tomographic screening for lung cancer. *N. Engl. J. Med.* **368**, 1980–1991 (2013).
11. Daskalakis, C. et al. Predictors of overall and test-specific colorectal cancer screening adherence. *Prev. Med.* **133**, 106022 (2020).
12. Parikh, N. D. et al. Biomarkers for the early detection of hepatocellular carcinoma. *Cancer Epidemiol. Biomark. Prev.* **29**, 2495–2503 (2020).
13. Stroun, M. et al. The origin and mechanism of circulating DNA. *Ann. N.Y. Acad. Sci.* **906**, 161–168 (2000).
14. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl Acad. Sci. USA* **112**, E5503–E5512 (2015).
15. Fece de la Cruz, F. & Corcoran, R. B. Methylation in cell-free DNA for early cancer detection. *Ann. Oncol.* **29**, 1351–1353 (2018).
16. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
17. Mathios, D. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).
18. Klein, E. A. et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).
19. Gao, Q. et al. Unintrusive multi-cancer detection by circulating cell-free DNA methylation sequencing (THUNDER): development and independent validation studies. *Ann. Oncol.* **34**, 486–495 (2023).
20. Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **11**, 3475 (2020).
21. Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
22. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
23. Bruhm, D. C. et al. Single-molecule genome-wide mutation profiles of cell-free DNA for non-invasive detection of cancer. *Nat. Genet.* **55**, 1301–1310 (2023).
24. Schrag, D. et al. Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. *Lancet* **402**, 1251–1260 (2023).
25. Nicholson, B. D. et al. Multi-cancer early detection test in symptomatic patients referred for cancer investigation in England and Wales (SYMPLIFY): a large-scale, observational cohort study. *Lancet Oncol.* **24**, 733–743 (2023).
26. Swanton C. et al. NHS-Galleri Trial Design: equitable study recruitment tactics for targeted population-level screening with a multi-cancer early detection (MCED) test. *J. Clin. Oncol.* **40**, TPS6606 (2022).
27. Hackshaw, A., Clarke, C. A. & Hartman, A.-R. New genomic technologies for multi-cancer early detection: rethinking the scope of cancer screening. *Cancer Cell* **40**, 109–113 (2022).
28. Trosman, J. R. et al. Perspectives of private payers on multicancer early-detection tests: informing research, implementation, and policy. *Health Aff. Scholar* **1**, qxad005 (2023).
29. Ma, X. et al. Multi-dimensional fragmentomic assay for ultrasensitive early detection of colorectal advanced adenoma and adenocarcinoma. *J. Hematol. Oncol.* **14**, 175 (2021).
30. Bao, H. et al. Letter to the Editor: An ultra-sensitive assay using cell-free DNA fragmentomics for multi-cancer early detection. *Mol. Cancer* **21**, 129 (2022).
31. Yu, P. et al. Multi-dimensional cell-free DNA-based liquid biopsy for sensitive early detection of gastric cancer. *Genome Med.* **16**, 79 (2024).
32. Zheng, R. et al. Cancer incidence and mortality in China, 2016. *J. Natl Cancer Cent.* **2**, 1–9 (2022).
33. Jonas, D. E. et al. *Screening for Lung Cancer With Low-Dose Computed Tomography: An Evidence Review for the U.S. Preventive Services Task Force* (Agency for Healthcare Research and Quality (US), 2021).
34. Bailey, S. E. et al. Diagnostic performance of a faecal immunochemical test for patients with low-risk symptoms of colorectal cancer in primary care: an evaluation in the South West of England. *Br. J. Cancer* **124**, 1231–1236 (2021).
35. Teoh, D. et al. Test performance of cervical cytology among adults with vs without human papillomavirus vaccination. *JAMA Netw. Open* **5**, e2214020 (2022).

36. Zeng, H. et al. Initial results from a multi-center population-based cluster randomized trial of esophageal and gastric cancer screening in China. *BMC Gastroenterol.* **20**, 398 (2020).

[1]Geneseeq Research Institute, Nanjing Geneseeq Technology Inc., Nanjing, China. [2]Nanjing Jiangbei New Area Center for Public Health Service, Nanjing, China. [3]The Fourth Affiliated Hospital of Nanjing Medical University, Nanjing, China. [4]Department of Thoracic Surgery and Oncology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. [5]Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai, China. [6]Colorectal Center, Jiangsu Cancer Hospital, Nanjing, China. [7]State Key Laboratory of Molecular Oncology, National Cancer Center, Cancer Hospital of the Chinese Academy of Medical Sciences, Beijing, China. [8]Department of Thoracic Surgery, Nanjing Drum Tower Hospital, Nanjing, China. [9]Department of Gynecologic Oncology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China. [10]Pancreas Center, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China. [11]Departments of Gynecology, Fujian Cancer Hospital and Fujian Medical University Cancer Hospital, Fujian, China. [12]Department of Clinical Laboratory, Guangdong Provincial People's Hospital, Guangzhou, China. [13]Department of Thoracic Surgery, Eastern Theater Command Hospital, Nanjing, China. [14]Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, Shanghai, China. [15]Department of Gynaecology, Second Xiangya Hospital of Central South University, Changsha, China. [16]Department of Gynecologic Oncology, Beijing Obstetrics and Gynecology Hospital, Beijing, China. [17]Department of Thoracic Surgery, The First Affiliated Hospital of China Medical University, Shenyang, China. [18]Department of Gastrointestinal Medical Oncology, Harbin Medical University Cancer Hospital, Harbin, China. [19]Department of Respiratory Medicine, First People's Hospital of Changzhou, Third Affiliated Hospital of Soochow University, Changzhou, China. [20]School of Public Health, Nanjing Medical University, Nanjing, China. [21]These authors contributed equally: Hua Bao, Shanshan Yang, Xiaoxi Chen. ✉e-mail: yang.shao@geneseeq.com

Methods

Study design and participants

The research design employed in the study was a multicenter, prospective, case–control approach aimed at detecting 13 types of cancers. The cancers targeted for detection included breast, cervical, colorectal, endometrial, esophageal, gastric, liver, lung, ovarian, pancreatic, prostate, bile duct and lymphoma. During the retrospective model training and validation phase, plasma samples were collected from 3,076 individuals diagnosed with one of the aforementioned cancer types. The collection period spanned from October 2017 to November 2020. In addition, plasma samples from 3,477 age-matched participants without cancer were also obtained. The MCED models were developed using the data from the training set. Subsequently, the models' performance was evaluated in the validation set to assess their predictive capabilities. The models were locked to ensure the integrity of the validation process.

In the subsequent, independent, prospective validation phase, patients who had received a pathological diagnosis of 1 of the 13 cancer types, as well as noncancer controls, were enrolled from April 2021 to November 2021 across multiple centers. Data analysts, who were responsible for generating and processing the sequencing data, remained unaware of the participants' clinical information. Conversely, those who collected the clinical data did not have access to the analytical results, maintaining a double-blind study design.

The conduct of the present study was under the oversight of the ethics committees of all the participating research centers, ensuring adherence to ethical standards. All individuals who took part in the study provided informed consent, having been made aware of the study's purpose and procedures. Detailed eligibility and exclusion criteria for participation in each phase of the study were meticulously outlined in Supplementary Note.

Sample collection, processing and sequencing

Retrospectively collected plasma samples were stored at −80°C and shipped to a centralized clinical testing center (Nanjing Geneseeq Technology Inc.; certified to College of American Pathologists, Clinical Laboratory Improvement Amendments and ISO15189) with dry ice. Prospectively collected blood samples, ~10 ml of each, were collected using EDTA tubes (Becton Dickinson) in each study site and transferred to a central laboratory (Nanjing Geneseeq Technology Inc.) at room temperature (preferably between 4 °C and 28 °C) for CANSCAN MCED testing. Blood samples were centrifuged at 16,000*g* for 10 min within 4 h of collection. Plasma cfDNA was extracted using the QIAamp Circulating Nucleic Acid Kit (QIAGEN) following the manufacturer's instructions. To ensure sufficient cfDNA for further analysis, the concentration of cfDNA in the plasma was determined using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) according to the manufacturer's guidelines.

The extraction of cfDNA was performed automatically on a Hamilton Microlab STAR, automated, liquid-handling platform. The cfDNA was extracted using magnetic-based bead extraction reagents included in the CANSCAN Sample Preparation Kit (Geneseeq) The cfDNA concentration was then measured with double-stranded DNA quantification reagents included in the sample preparation kit. The amount and the purity of extracted DNA samples were measured to meet the minimum required standards of the test. After the extracted nucleic acids had been quantified by cfDNA, the total amount of nucleic acids should not be <l5 ng to pass sample quality control. If the concentration of the extracted nucleic acid was <0.4 ng μl^{-1}, it should be dried and concentrated at <45 °C to reach or exceed a concentration of 0.4 ng μl^{-1}.

Then, 5–10 ng of cfDNA per sample was subjected to PCR-free WGS library construction using the library construction reagents included in the CANSCAN Sample Preparation Kit. The library was constructed automatically on Biomek, quantified using the quantitative (q)PCR reagents included in the CANSCAN Sample Preparation Kit and underwent 150-bp, paired-end sequencing on NovaSeq 6000 platforms. Sample libraries were quantified and normalized into a sequencing pool. Pooled sample libraries were fluorometrically quantified, loaded on a sequencing flow cell and sequenced on NovaSeq 6000 platforms. Clinical negative controls were involved to ensure that there was no DNA contamination during the library preparation.

To ensure data quality, the sequencing output was subjected to quality control measures. Trimmomatic[37] was used for read trimming, followed by PCR duplicate removal using Picard tools (Broad Institute). The trimmed reads were aligned to the human reference genome (GRCh37/UCSC hg19) using the Burrows–Wheeler Aligner[38]. To standardize the data and mitigate the effects of variable sequencing depth, we applied a downsampling procedure. Samples should not be <5× average coverage depths to pass the sequencing quality control. Coverage depths that exceeded 5× were reduced to a uniform 5×, ensuring consistency across these samples.

Jinling prospective cohort study

The Jinling study is a prospective, multicenter, observational cohort designed to enroll 15,000 individuals aged 45–75 years at two physical examination sites. Participants underwent annual routine physical examinations and MCED testing for 3 years, followed by a 2-year active and passive follow-up period. Participants were excluded if they had a history of cancer or current diagnosis of cancer or had received an organ transplant or prior nonautologous (allogeneic) bone marrow or stem-cell transplant. Participants provided written informed consent. The protocol, including the interim analysis, was approved by the Nanjing Jiangbei Hospital Ethics Committee and the Fourth Affiliated Hospital of Nanjing Medical University Ethics Committee (protocol nos. XZ-2021032 and 20230106-k076). This ensured that all ethical considerations were thoroughly evaluated, with approval numbers and statements confirming compliance with the Declaration of Helsinki and relevant regulations. The study was registered with ClinicalTrials.gov, identifier NCT06011694. Beyond the primary study components of physical examinations for 3 years, participants were also provided information that might indicate increased hereditary risk to selected cancer types, derived from separate genetic testing on collected white blood cells, constituting an additional incentive without charge to the participants.

Participants in the study were enrolled in annual routine physical examinations. At the time of attending this annual routine physical examination, participants also contributed ~10 ml of blood into an EDTA cfDNA tube. Samples were couriered to a central laboratory (Nanjing Geneseeq Technology Inc.). Plasma isolation, followed by cfDNA extraction, was carried out as previously mentioned. Participant data were collected, including self-reported demographics (for example, age and sex), lifestyle factors (for example, smoking status, alcohol use, dietary habits, living environment) and health history (for example, family history of cancer-related disease history, history of other related diseases). Sex was determined based on both self-reporting at birth and WGS data, with no discrepancy observed. Participants were not included in the study if they withdrew consent or their blood samples or extracted cfDNA did not pass quality control. On completion of the routine physical examinations, participants would proceed with cancer clinical evaluations by their physicians, based on the individual health profiles and physical examination outcomes. All participants received proper treatment based on current standard-of-care procedures after their physical exams. The Jinling study employed a clinical research organization (Suzhou Brilliancy Health, Inc.) to maintain data integrity and objectivity throughout the investigation. In line with these measures to ensure rigor, the results of the MCED test were deliberately withheld from both the participants and their healthcare providers. This blinding procedure was implemented to prevent the study's findings from influencing clinical care decisions and to mitigate potential bias. Specifically, the MCED test (CANSCAN, Geneseeq) was conducted without any prior knowledge of the participants' clinical outcomes.

Follow-up procedures for the Jinling study were conducted through two primary methods: active and passive follow-up. Data were to be collected within 3 months of enrollment to identify cancer diagnoses and serious disease outcomes. Active follow-up was performed every 3 months for those without a diagnosis at 3 months and involved direct contact with participants through follow-up questionnaires by text message or telephone calls. This proactive approach aimed to gather up-to-date information on the participants' health status, any new diagnoses of cancer, treatments received and overall well-being. The passive follow-up component entailed a retrospective review of electronic health records every 6 months, which was conducted with the explicit informed consent of the participants. This consent authorized the researchers to access and analyze their medical data, as stored within the Jiangsu Commission of Health's healthcare system.

Cancer diagnosis was determined using the classification system of *International Classification of Diseases for Oncology* (ICD-O)[39]. This involves pathological confirmation of an invasive solid or hematological malignancy, with ICD-O codes of /3, /6 or /9, as well as premalignant conditions with an elevated risk of malignancy indicated by an ICD-O code of /2 (for example, villous adenoma of colorectum, high-grade squamous intraepithelial lesion). If a participant was diagnosed with or died from cancer or one of the premalignant conditions defined above, they would be categorized in the cancer diagnosis group. Conversely, other premalignant conditions (for example, tubular adenoma of the colorectum, cervical intraepithelial neoplasia grade 1, gastric polyps) were not considered to be cancer.

For this interim analysis, we considered participants who enrolled in the first year of study. All participants were followed up for at least 1 year. The data for the interim analysis were extracted from the initial physical examination records, MCED test results and findings of their follow-up clinical workups. All data were anonymously coded and stored in a secure database in accordance with ethics guidelines. Results from participants' physical examinations were collected, with a particular emphasis on cancer-related abnormalities. The cancer screenings conducted as part of these examinations included those for colorectal, cervical, prostate and lung cancers, which align with USPSTF grades A, B or C, as well as screenings conducted for cancers that do not currently have USPSTF recommendations, specifically those of the breast, liver, kidney, thyroid and pancreas. More detail on the definitions of cancer-related abnormalities is provided in the appendix study protocol ('Jinling Cohort Protocol' in Supplementary Information).

Multicancer detection classifier and tissue of origin prediction

A customized analytical framework was devised using a multilayer stacked ensemble methodology. This framework processed WGS data to extract a suite of multidimensional genomic and fragmentomic features. The feature frameworks included copy number variations (CNVs), fragment size coverage (FSC), fragment size distribution (FSD), nucleosome footprint (NF) and fragment-based methylation (FM). Each feature framework was independently analyzed to generate foundational data. Subsequently, for each feature framework, one binary base model aimed to detect cancer-specific signals and one multiclass base model designed to predict the TOO for these signals was developed. The binary base model, tasked with distinguishing between cancerous and noncancerous signals, was constructed using neural network architectures. The multiclass base model, which was responsible for predicting the TOO to one of the predetermined cancer sites of the detected signals, employed a stacked ensemble approach combining the strengths of Extreme Gradient Boosting (XGBoost), generalized linear models (GLMs) and deep learning algorithms.

CNV model. The profiling of CNVs denotes the variation in the number of copies of DNA segments in the human genome. Specifically, a CNV represents an increase (amplification) or decrease (deletion) of DNA segments that are at least 1 kb in size. These variations form an important component of the genetic structural variation landscape. To analyze CNVs, the genome of each sample is first segmented into 1-Mb bins. For each of these bins, the depth after bin-level GC correction is assessed through a hidden Markov model. This model compares the sample's depth against a baseline derived from healthy samples for the same bin. Subsequently, the $\log_2$(ratio) between the sample and the baseline is calculated for each bin, providing insights into the CNV patterns.

To model CNV features for the binary base model, a convolutional neural network (CNN) was employed because of its proficiency in extracting spatial hierarchies and patterns from data. This CNN architecture layered convolutional blocks consisting of convolutional layers, ReLU (rectified linear unit) activation, batch normalization and pooling layers. These blocks sequentially processed the 1-Mb bin data, capturing and abstracting the key features indicative of CNVs. After feature extraction through the convolutional blocks, a fully connected layer classified the CNV patterns detected. This combination offered a nuanced approach to identifying and classifying CNV features, making critical contributions to comprehending the genetic structural variation landscape.

FSC model. The FSC was generated by comparing the coverages of short (65–150 bp) and long (150–220 bp) fragments. The extraction of fragment size features was adapted from the method introduced by Mathios et al.[17]. A nonparametric method for fragment-level adjustment for GC content and library size was first performed. The genome was first divided into 100-kb bins. Next, the coverage of the two fragment size groups in each 100-kb bin was calculated and corrected by the GC content. The coverages in every 50 continuous 100-kb bins were combined to calculate the coverage in the corresponding 5-Mb window, resulting in 541 windows (windows with <50 bins were discarded). For each fragmentation size group (short and long), the scaled coverage score (z-score) in every 5-Mb bin was calculated.

To analyze the FSC data, a similar CNN structure was employed. By processing the z-score sequences from the 5-Mb windows, the CNN adeptly identified features associated with fragmentation patterns, offering nuanced insights into genomic structure variations driven by fragment size disparities.

FSD model. The FSD feature examined the coverage of cfDNA fragments ranging from 100 bp to 220 bp for every 5 bp (for example, 110–114 bps, 115–119 bps, …, 215–220 bps; 24 bins) at chromosome arm level. This granular approach captures the nuances of cfDNA fragmentation patterns that reflect cancer-associated nuclease activities and chromatin organization. Cancer-related cfDNA often exhibits distinct fragmentation patterns as a result of increased cell turnover, apoptosis and altered nuclease activities. Notably, peaks slightly below 100 bp and around 200 bp corresponded to mononucleosomal and dinucleosomal fragments, respectively, as previously described by Jahr et al.[40]. Variations in the height and definition of these peaks among different sample groups provided insights into the physiological and pathological states affecting cfDNA characteristics.

In total, 39 chromosome arms were examined, generating 936 (39 arms × 24 bins) FSD features. The short arms of five acrocentric chromosomes were not included because they remained largely unsequenced to date. The raw coverage of FSD was also scaled into the z-score by comparing the variable value against the overall mean value to ensure that the model focused on the underlying patterns in the data, rather than being influenced by sample-to-sample differences. This approach enabled the enhanced detection of high-resolution chromosome-level patterns, which could potentially reveal further distinctions between the cancer and noncancer groups.

A similar CNN structure was used to systematically analyze the FSD features. By processing the z-score-transformed sequences corresponding to the stepwise cfDNA fragment lengths, the CNN

effectively discerned patterns indicative of chromosomal aberrations. This method underscored the potential of employing a uniform CNN architecture to excavate deep insights from genomic data variations, particularly in distinguishing pathological states based on fragment size distributions at the chromosome level.

NP model. NP features captured the occupancy and positioning of nucleosomes across the genome, which were indicative of chromatin accessibility and transcriptional activity[41]. In cfDNA, nucleosome footprints could reflect the epigenetic landscape of the cells of origin and had been well demonstrated for cancer detection and tumor TOO prediction in several studies[22,42–44]. At actively bound transcription factor-binding sites (TFBSs), nucleosome displacement allowed DNA-binding proteins to access the DNA, resulting in characteristic fragmentation patterns in cfDNA sequencing data.

Our NP framework, adapted from methodologies, described by Doebley et al.[45], assesses nucleosome occupancy around selected TFBSs that are highly correlated with cancer processes. By analyzing the protected (nucleosome-bound) and unprotected regions, we could infer changes in chromatin structure associated with tumorigenesis. Briefly, a set of 854 TFBSs was selected. All reads in a window of −5,000 bp to +5,000 bp around each site with fragment length of 100–220 bp were extracted. Sites were then spited into 15-bp bins and GC corrected based on the weighted fragment midpoints in each bin. Next, bins within regions with known mapping problems or extremely high coverage were excluded. Finally, bins generated for all sites were normalized to generate the mean coverage profiles for all sites. To quantify the coverage profiles, for each TFBS, three values were calculated based on the filtered GC-corrected coverage profile: (1) central coverage, measured by coverage value at a distance of 30 bp from the specified binding site; (2) average coverage, calculated by the average of coverage of ±1,000-bp regions from the binding site; and (3) amplitude of the binding site, ascertained through the application of a fast Fourier transform analysis. In addition, we incorporated orientation-aware cfDNA fragmentation values[46], referencing open chromatin data from a T cell line (GSM665839). This T cell line is representative of the hematopoietic system's contribution to the ctDNA pool. Notably, these values tend to decrease in cancer patients, reflecting alterations in the hematopoietic system's DNA contribution[46].

A similar CNN structure was used to systematically analyze the NP features. This network structure effectively processed the input features—central coverage, average coverage and amplitude—to classify samples and distinguish cancer presence from low-pass WGS data, illustrating the power of applying advanced neural network models to genomic data analysis.

FM model. FM provides an additional layer of information by examining the cleavage patterns around cytosine phosphate–guanine (CpG) sites to infer the methylation status of samples. An increased likelihood of cfDNA cleavage at the adjacent cytosine is associated with a methylated CpG site, as described by Zhou et al.[47]. This approach leveraged the pattern of cfDNA cleavage to inform the methylation levels of genomic regions of interest without requiring bisulfite conversion. We focused on Alu repetitive elements because pronounced methylation levels are observed in these regions compared with the overall human genome and standard CpG islands. Alu elements also display different proportions of methylation in cancer versus normal cells and among different cancer types[48,49].

Within these Alu regions, the fragment ratio of the 5′-end motifs of CGN and NCG (with N representing any nucleotide: A, C, T or G) was analyzed, resulting in eight distinct fragment ratios being identified. The ratios of CGN:NCG motifs were also calculated. A notably elevated cleavage rate at both C positions was found in the CGCG sequencing. In a broader genomic context, all two-tandem CpG nucleotides spanning the CGCG sequence were examined. The fragment ratios derived from the 5′-end motifs of CGC, NCG(C1), CGN and CGC(C2) were identified as features. In total, ten fragment ratios for CGCG sequencing sites were documented. In addition, the motif ratios of CGC:NCG and CGN:CGC were calculated and documented as feature sets. A fully connected neural (FCN) network was deployed for the FM features. This architectural choice is suited to handling the structured data generated from the profiling framework, allowing the network to learn and identify complex patterns indicative of cancer. The FCN network comprises multiple dense layers, with each layer connected to its predecessor, and incorporates nonlinear activation functions between layers to model the intricate relationships between nucleosome occupancy patterns and cancer signal.

Ensemble generalized linear classifiers

The output generated by these base models was then directed into two separate ensembles composed of GLMs. For the binary classification ensemble tasked with detecting cancer signals, logistic regression models the probability of an instance being classified as 'cancer' versus 'no cancer'. For the TOO ensemble, the multinomial logistic regression model extends this to multiple classes. The probability that an observation $\mathbf{x}$ belongs to a specific class c among K possible classes is given by:

$$\hat{y}_c = \Pr(y = c|\mathbf{x}) = \frac{e^{\mathbf{x}\,\boldsymbol{\beta}_c}}{\sum_{k=1}^{K} e^{\mathbf{x}\,\boldsymbol{\beta}_k}}$$

where, $\mathbf{x}$ denotes the transpose of the input feature vector $\mathbf{x}$, and $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_k$ are the coefficient vectors for classes c and k, respectively. The denominator is the sum of the exponentiated model linear predictors across all K classes, ensuring that the probabilities sum up to 1.

Thresholds for decision calls. During the training of the classifiers, internal validation was leveraged to establish thresholds that would correspond to predetermined specificity levels. For the binary classification ensemble aimed at cancer detection, two thresholds were defined based on the validation set. The SST was set such that the specificity of the model reached 99.0%. Specificity is the ability of the test to correctly identify noncancer cases (true negatives) among all the noncancer cases that it evaluates. A specificity of 99.0% means that 99 out of 100 noncancer cases are correctly identified as such by the model. The HST was set at a higher specificity level of 99.9%. This more stringent criterion ensures that the model is even more conservative in labeling a case as 'cancer', thus reducing the FPR further.

For TOO classifications, a different approach was taken. Instead of using a single threshold, the decision was based on the difference between the highest and the second-highest scores predicted by the model for the various tissues of origin. This method, often referred to as the 'top two TOO score difference', provides a measure of confidence in the model's prediction. A large difference between the top two scores indicates a higher confidence in the classification, because the model shows a clear preference for one tissue over the others.

By utilizing these threshold strategies, the classifiers can be fine-tuned to meet specific performance metrics, particularly in terms of reducing false positives, which is crucial in medical diagnostics.

Statistical analyses

All statistical analyses were carried out using Python (3.10.9 or 3.8.0) and R (4.2.3). Machine learning models were built using PyTorch[50] (v.2.2.1) and the H2O (ref. 51) (v.3.42.0.2) framework and continuous variables were summarized using the median and IQR, whereas categorical variables were described using counts and percentages. The CIs for sensitivity, specificity, PPV and NPV were calculated using Wilson's score interval formula. The significance threshold was set at 5%, with all 95% CIs being two sided, unless otherwise specified. Pre-test probabilities were based on the prevalence of cancer reported in the Jinling study for each subgroup, defined by the presence or absence

of cancer-related abnormalities found during physical examinations. Additional analyses for the Jinling cohort have been described in the statistical analysis plan of the study protocol of Jinling cohort ('Jinling Cohort Protocol' in Supplementary Information).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The reference human genome used for analysis is GRCh37 (d5; https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence). The study protocol and statistical analysis plan are available within Supplementary Information. All feature datasets analyzed for the present study are available via the Geneseeq research data portal at https://zh.geneseeq.com/mercury/datause-license-request.html. Approved researchers can access and download these datasets following registration and acceptance of a data-sharing agreement on the portal.

Code availability

All code used for reported analyses has been made available through the Geneseeq research data portal, available at https://zh.geneseeq.com/mercury/datause-license-request.html.

References

37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Fritz, A. et al. *International Classification of Diseases for Oncology* 3rd edn (World Health Organization, 2000).
40. Jahr, S. et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **61**, 1659–1665 (2001).
41. Lai, B. et al. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**, 281–285 (2018).
42. Peneder, P. et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat. Commun.* **12**, 3230 (2021).
43. Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
44. Markus, H. et al. Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Sci. Transl. Med.* **13**, eaaz3088 (2021).
45. Doebley, A.-L. et al. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. *Nat. Commun.* **13**, 7475 (2022).
46. Sun, K. et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
47. Zhou, Q. et al. Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc. Natl Acad. Sci. USA* **119**, e2209852119 (2022).
48. Joyce, B. T. et al. Prospective changes in global DNA methylation and cancer incidence and mortality. *Br. J. Cancer* **115**, 465–472 (2016).
49. Rodriguez, J. et al. Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. *Nucleic Acids Res.* **36**, 770–784 (2007).
50. Ansel, J. et al. PyTorch 2: faster machine learning through dynamic Python bytecode transformation and graph compilation. In *Proc. 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* 929–947 (Association for Computing Machinery, 2024).
51. LeDell, E. & Poirier, S. H2O AutoML: scalable automatic machine learning. In *Proc. AutoML Workshop at ICML* 24 (ICML, 2020).

Acknowledgements

We thank all the patients, study participants and their families for their contributions to this effort. We thank the investigators and research staff at all study sites. We also thank the National Healthcare Big Data (East) Center-Nanjing for assisting with the Jinling cohort follow-up by providing the healthcare database. The present study was sponsored by Nanjing Geneseeq Technology Inc. We thank A. Margaritescu and L. Hummel for their assistance with language editing.

Author contributions

Conceptualization: H.B., S.Y., X. Zhang, Xue Wu and Y. Shao. Methodology: H.B., S.Y., X. Chen, X. Cheng, S.W., M.W., S.T., S.C., P.H., X. Xu, J.Z., Y. Shen, Y.J., S.L. and Y. Shao. Software: X. Chen, X. Cheng, M.W., S.T., S.C., P.H., X. Xu, J.Z. and Y. Shen. Validation: H.B., S.Y., X. Chen, Xiuxiu Xu and S.L. Formal analysis: H.B., S.Y., X. Chen and X. Cheng. Investigation: H.B., S.Y., X. Chen, G.D., Y.M. and Y. Shao. Visualization: X. Chen, Xiuxiu Xu and W.T. Resources: G.D., Y.M., Z.W., L.Y., J.L., T.W., B.Z., K.J., Q.X., J.C., H.H., J.P., X. Xia, Y.W., S.X., J.T., L.C. and D.Z. Data curation: X. Cheng, D.Z. and R.Y. Writing—original draft: H.B., S.Y. and X. Chen. Writing––review and editing: H.B., S.Y., X. Chen, X. Cheng, Xiuxiu Xu, W.T., S.L. and Y. Shao. Project administration: H.B., S.Y., Xue Wu, G.D., Y.M. and D.Z. Supervision: Xue Wu, X. Wang, X. Zhang and Y. Shao.

Competing interests

H.B., S.Y., X. Chen, S.W., X. Cheng, W.T., M.W., S.T., D.Z., R.Y., S.C., P.H., X. Xu, J.Z., Y. Shen, Y.J., S.L., X.Z., Xue Wu, X. Wang and Y. Shao are employees of Nanjing Geneseeq Technology Inc. The other authors declare no competing interests.

Additional information

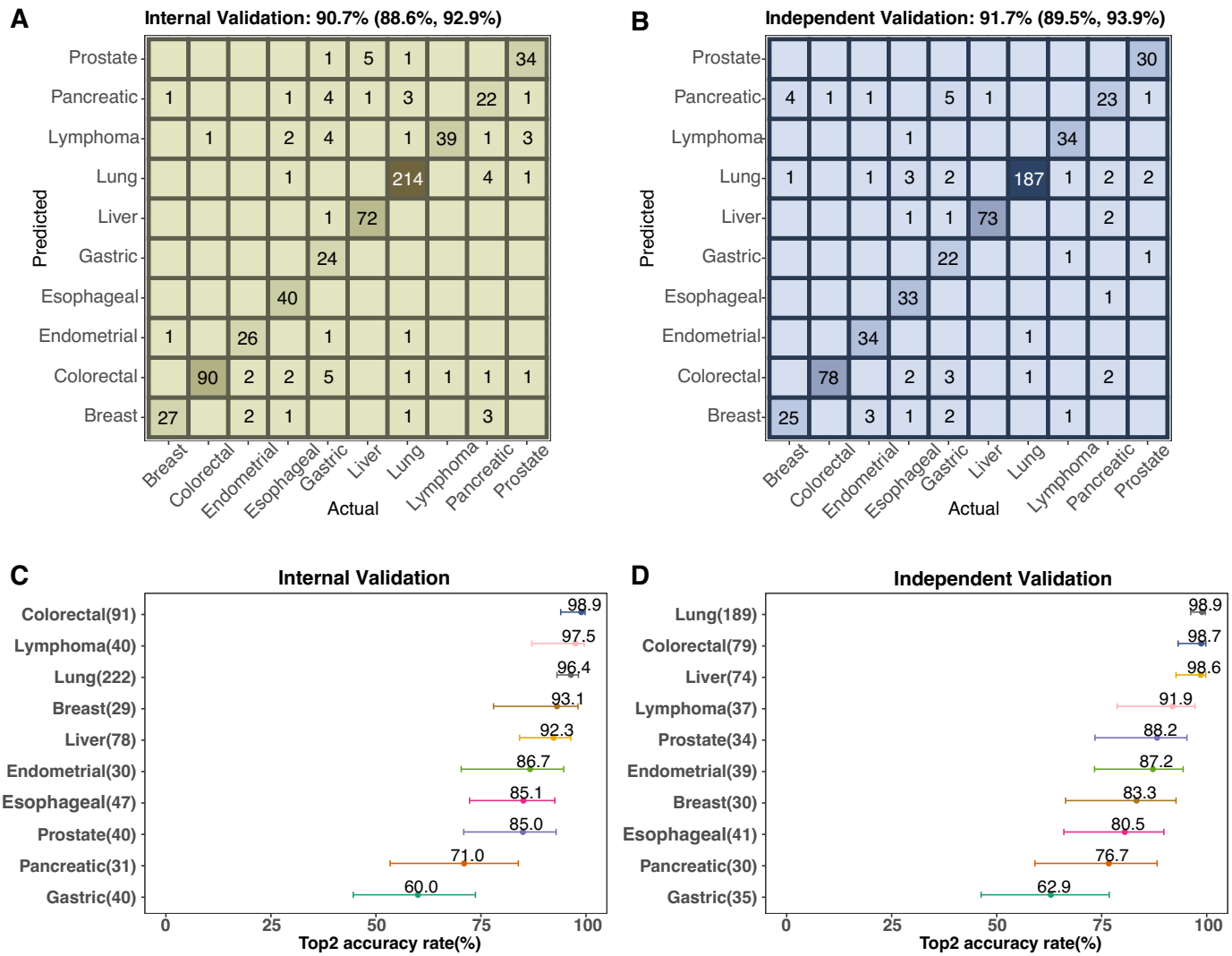
Extended data is available for this paper at https://doi.org/10.1038/s41591-025-03735-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-03735-2.

Correspondence and requests for materials should be addressed to Yang Shao.

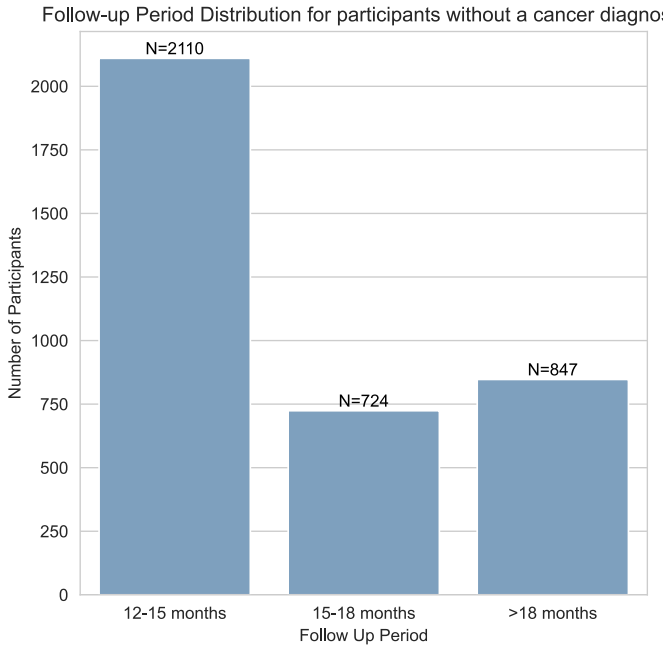
Peer review information *Nature Medicine* thanks Ziding Feng, Dominic Rothwell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Ulrike Harjes, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-025-03735-2



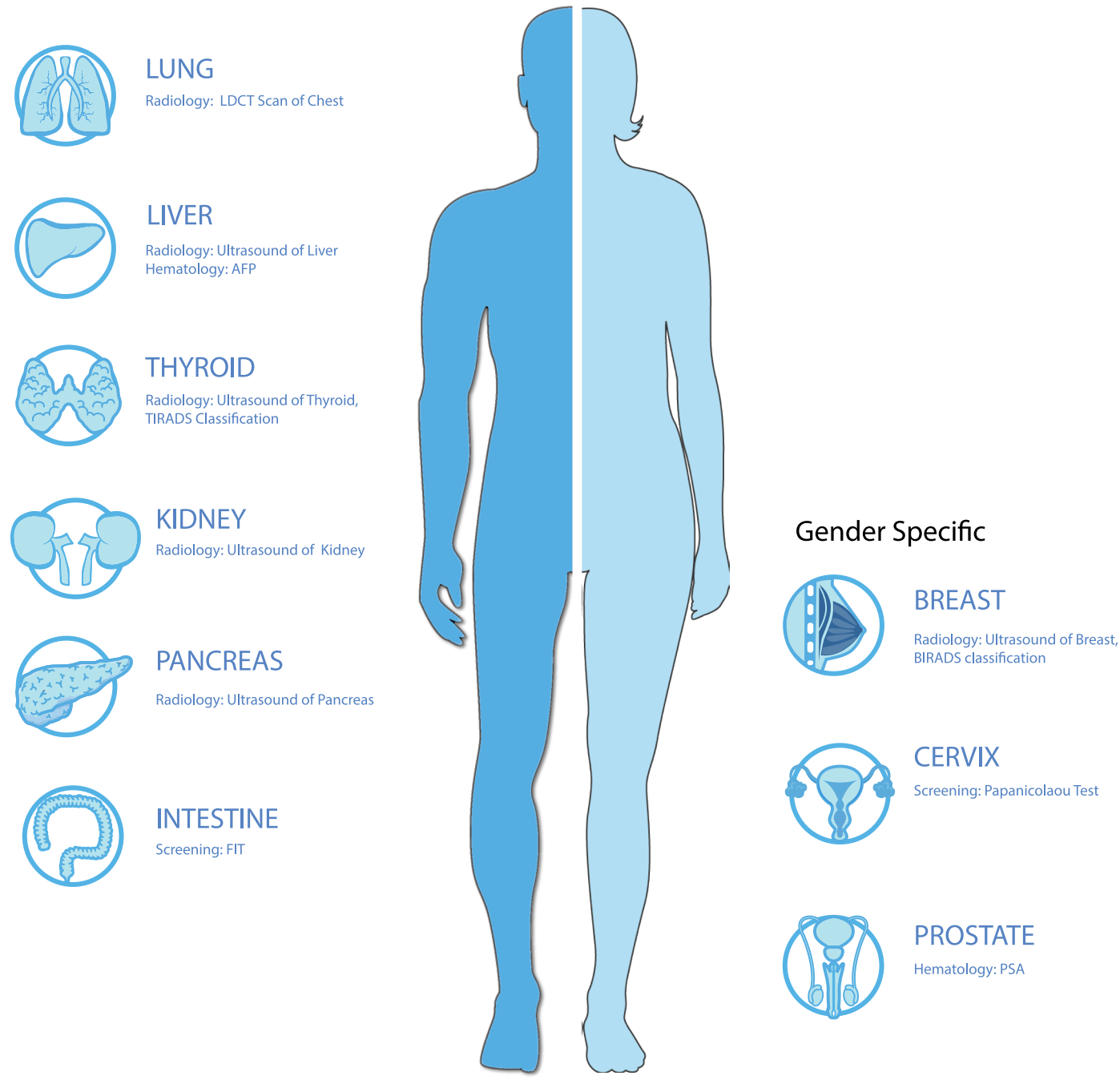
Extended Data Fig. 1 | Top 2 Tissue of Origin prediction (TOP2) accuracy for ten cancer types. (**a**) Confusion matrix displaying the predicted and actual cancer types for the internal validation set. (**b**) Confusion matrix displaying the predicted and actual cancer types for the independent validation set. (**c**) TOP2 accuracy rates for different cancer types in the internal validation set, with error bars representing the 95% confidence interval of accuracy. The number of patients in each cancer type is indicated in parentheses. (**d**) TOP2 accuracy rates for different cancer types in the independent validation set, with error bars representing the 95% confidence interval of accuracy. The number of patients in each cancer type is indicated in parentheses. This analysis is based on independent patient samples, with the unit of study being the individual patient. No technical or biological replicates were performed at the individual patient sample level for this analysis; the 'n' values represent counts of unique patients.

https://doi.org/10.1038/s41591-025-03735-2



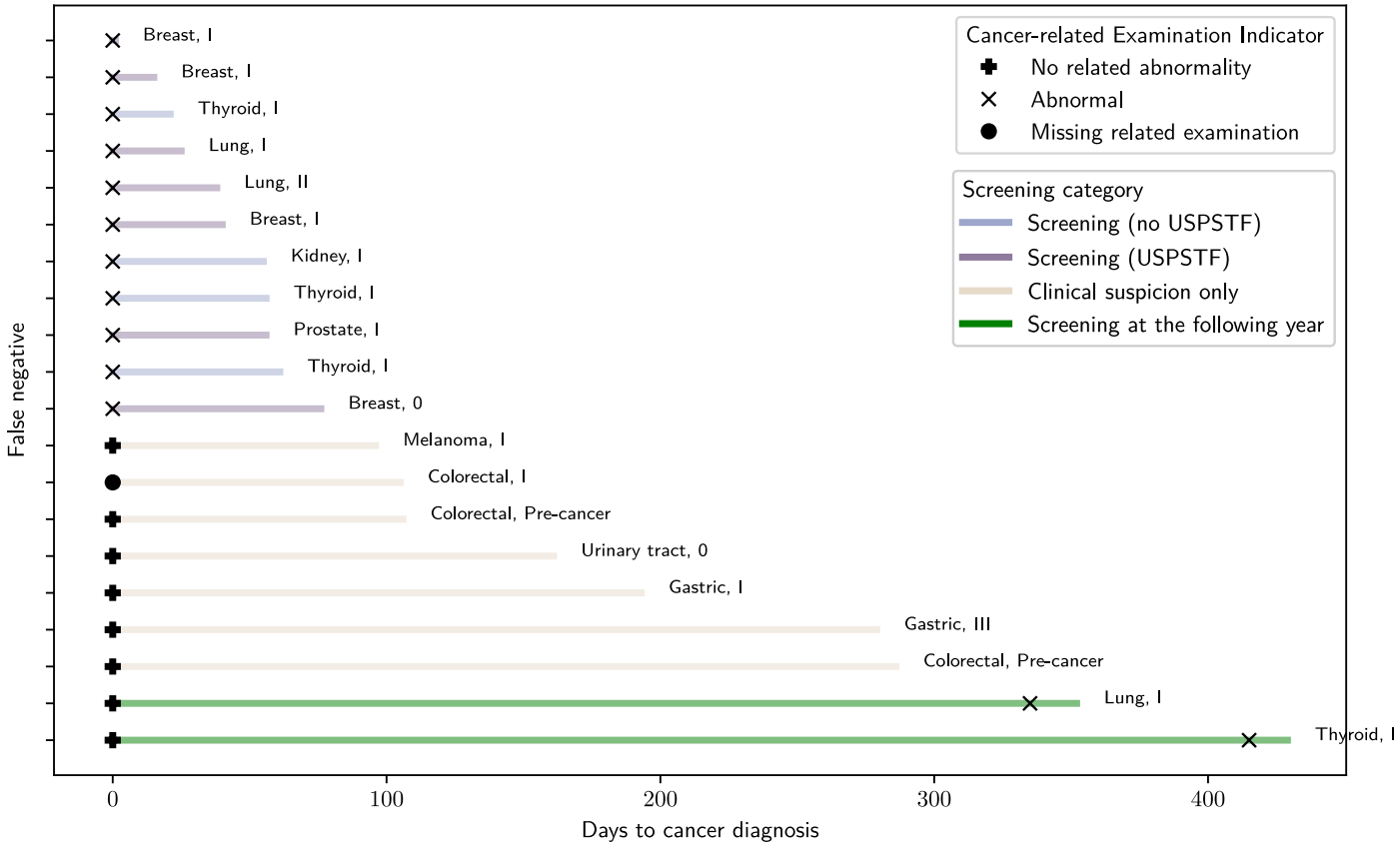
Extended Data Fig. 2 | Follow-up period distribution for participants without a cancer diagnosis. The majority of participants (N=2,110) were followed for 12–15 months, while 724 participants were followed for 15–18 months, and 847 participants were followed for over 18 months.

https://doi.org/10.1038/s41591-025-03735-2



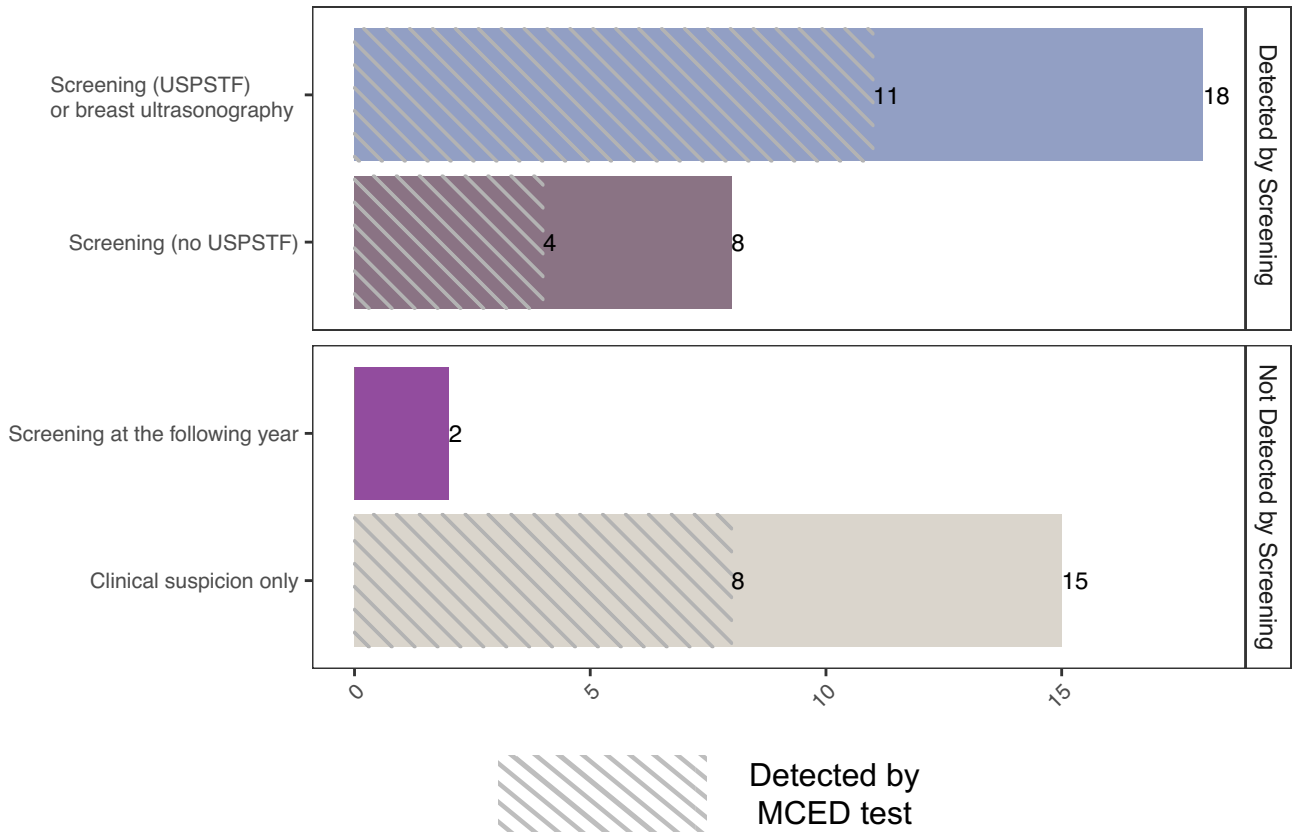
Extended Data Fig. 3 | Cancer abnormality-related physical examination procedures. The physical examination procedures employed in the study. Procedures applied to all participants included assessments of the lungs, liver, thyroid, kidneys, pancreas, and intestines. Gender-specific procedures encompassed evaluations of the breasts, cervix, and prostate.

https://doi.org/10.1038/s41591-025-03735-2



Extended Data Fig. 4 | Time to cancer diagnosis and cancer-related examination indicators for false negative MCED test participants. Days to cancer diagnosis for each false negative case.

https://doi.org/10.1038/s41591-025-03735-2



Extended Data Fig. 5 | Cancer detection by the MCED test compared to existing screening methods. The concordance between cancer cases identified by the MCED test and those detected through existing screening protocols. Screening (USPSTF) or breast ultrasonography: cancers detected by both the MCED test and screening methods recommended by the United States Preventive Services Task Force (USPSTF) or breast ultrasonography. Screening (no USPSTF): cancers detected by both the MCED test and screening methods not currently recommended by USPSTF. Screening at the following year: cancers detected by the MCED test during screening at the following year's appointment. Clinical suspicion only: cancers detected by the MCED test based on clinical suspicion only.

nature portfolio

Corresponding author(s): Yang Shao

Last updated by author(s): 14/5/25

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	No software was used for data collection
Data analysis	All statistical analyses were carried out using python (3.10.9 or 3.8.0) and R (4.2.3). Trimmomatic v0.32 was used for FASTQ file quality control. BWA v0.7.12 was used for mapping sequencing reads to the reference genome. Picard v2.27.0 was used for removing PCR duplicates. MutationalPattern v1.10.0 was used for fitting mutational signatures into 96 context profiles. Machine learning model were built using pytorch(2.2.1) and H2O(v3.42.0.2) framework.

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 reference human genome used for analysis is GRCh37 (d5; https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/

phase2_reference_assembly_sequence/). The study protocol and statistical analysis plan are available with the publication in the supplementary information. All feature datasets analysed for this study are accessible via the Geneseeq Research Data Portal at https://zh.geneseeq.com/mercury/datause-license-request.html. Approved researchers can access and download these datasets following registration and acceptance of a data sharing agreement on the portal.

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	Our findings apply to both sexes. Participants were matched for age and sex in the study design. Sex was determined based on both self-reporting at birth and WGS data, with no discrepancy observed. Consent has been obtained for sharing data. Participant sex (with no gender data collected), was considered in the study design and analysis, in the context of cancer incidence rates.
Reporting on race, ethnicity, or other socially relevant groupings	No socially constructed or socially relevant categorization variable was used in the study.
Population characteristics	The study population for this research consisted of a multi-cohort design incorporating training, internal validation, and independent validation sets to rigorously assess the research aims. The training cohort was the largest, comprising 4,807 participants, divided into 2,254 cancer patients (1,108 females and 1,146 males) and 2,553 non-cancer controls. Cancer patients in this training set exhibited a median age of 59 years (Interquartile Range [IQR]: 51-67 years) while non-cancer participants showed a slightly younger median age of 56 years (IQR: 52-62 years). An internal validation cohort (n=1,746), composed of 822 cancer patients and 924 non-cancer individuals, displayed comparable age characteristics, with cancer patients presenting a median age of 59 years (IQR: 52-68 years) and non-cancer participants a median age of 56 years (IQR: 52-62 years). To ensure the generalizability of findings, an independent validation cohort (n=1,465) was included, with an even split between 732 cancer patients and 733 non-cancer participants. The age profiles were again consistent, with cancer patient median age at 59 years (IQR: 51-67 years) and non-cancer participants at 57 years (IQR: 52-62 years). The cancer cohorts across all three datasets were characterized by diagnostic heterogeneity, encompassing 13 distinct cancer types. The age distributions and comparability between cancer and non-cancer groups within each cohort will be a factor considered in subsequent analyses. Additionally, data from the ongoing Jinling study, which included 3,724 asymptomatic participants after applying exclusion criteria, was also available for analyses. This cohort was predominantly female (65.4%, n=2435) with the remaining 34.6% being male participants (n=1289). The age distribution ranged different age groups, the major groups being, 51-55 (31.1%), 56-60 (28.8%) and 45-50 (14.8%). The majority of participants in Jinling Study were non-smokers(79.5%). The age and demographic distributions, including the Jinling cohort's characteristics, were considered factors in subsequent analyses to ensure the validity and generalizability of the study's findings.
Recruitment	In the model construction phase and independent validation phase, participants were recruited from multiple participating research centers. For Jinling cohort study, Residents from local communities in the Jiangbei New Area of Nanjing will be invited to participate. The combined effects of self-selection bias within the Jinling cohort, geographic bias from recruiting solely in Nanjing and east asian population.While the model might perform well within the specific context of the Jinling cohort in Nanjing, its utility and robustness for broader cancer patient populations and different geographic regions. Acknowledging and explicitly discussing these biases as study limitations is crucial for responsible interpretation and application of the research findings. Future efforts should consider strategies to diversify recruitment, expand the scope of cancer types studied, or conduct external validation in geographically and demographically distinct populations to assess and improve the model's generalizability and real-world impact.
Ethics oversight	The conduct of this study was under the oversight of the Ethics Committees of all participating research centers, ensuring adherence to ethical standards. For Jinling Cohort study, the protocol and informed consent, including the interim analysis, was approved by the Nanjing Jiangbei Hospital Ethics Committee and the Fourth Affiliated Hospital of Nanjing Medical University Ethics Committee(XZ-2021032, 20230106-k076).

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	The determination of the participant sample size for this study is predicated on the prevalence of cancer in China, coupled with the anticipated sensitivity and specificity of the Multicancer Early Detection (MCED) test under evaluation. Employing a single-group target value methodology, the sample size calculation aims to rigorously validate the performance characteristics of the CanScan MCED technology.

Preliminary data suggest an expected sensitivity of 85%. To ensure adequate statistical power to detect a clinically meaningful difference, a target sensitivity difference of 15% was established, resulting in a target sensitivity value of 70%. With a prespecified alpha (α) level of 0.025 (one-sided) and a desired power (1-β) of 0.8, the calculation dictates a need for observing 64 cancer incidences within the study cohort. Considering the reported five-year cumulative cancer incidence rate of 0.55% in the target population and accounting for an anticipated follow-up attrition rate of 20%, a total of 14,517 participants is required to yield the necessary number of cancer events for robust sensitivity assessment.

Similarly, to evaluate specificity, an expected value of 95% is posited for the CanScan MCED technology. Given the importance of minimizing false-positive results and ensuring clinical utility, a target specificity difference of 5% was deemed appropriate, establishing a target specificity value at 90%. Utilizing the same statistical parameters (α = 0.025, one-sided; 1-β = 0.8), this specificity assessment necessitates the inclusion of 239 healthy individuals. Consequently, the combined requirements for sensitivity and specificity validation point to an overall necessary sample size.

Considering the substantially larger sample required for sensitivity assessment due to the lower incidence of cancer, the definitive sample size for this project is conservatively set at approximately 15,000 qualified participants. This enlarged cohort is designed to comprehensively fulfill the statistical requirements not only for both sensitivity and specificity validation but, crucially, to ensure sufficient cancer events are observed to robustly assess the sensitivity of the CanScan MCED technology in a real-world context.

Data exclusions

The numbers of participants described in the study were after exclusions based on analysis protocal. No other data was excluded from analyses.

Replication

This study is a cohort study and therefore is not applicable for experimental replication.

Randomization

This study did not employ randomization. Participants were selected based on their specific cancer types/healthy status and were not randomly assigned to treatment groups. Specifically, the Jinling study is designed as an observational cohort, not a randomized controlled trial (RCT); therefore, no randomization procedure was implemented.

Blinding

For case-control study part, investigators responsible for plasma sample collection and processing were rigorously blinded to diagnostic results to mitigate potential ascertainment bias. Blinding was not applicable for subsequent analyses explicitly focused on comparisons between pre-defined participant groups based on diagnostic classifications. Once participants are categorized into groups (e.g., comparing groups with different cancer stages or types), the analytical objective shifts to examining differences between these known and defined groups. At this analytic stage, blinding becomes not only logistically impractical but also conceptually irrelevant, as the very nature of the comparison relies on the distinct and known group memberships for the diagnostic categories under investigation. The purpose of these analyses is to quantify and characterize the differences, if any, between these diagnostically defined groups, necessitating the utilization and acknowledgement of group labels derived from the very diagnostic results that were blinded to prior to this stage.
An independent CRO managed de-identification and unblinding producers to ensure data integrity and objectivity throughout the Jinling cohort study. The CRO performed the sample blinding process, assigning unique, de-identified sample IDs. These blinded IDs were then provided to Geneseeq to label the plasma samples for sequencing and subsequent analyses. For interim analysis and eventual study readout, the CRO provided the unblinding key to Geneseeq, solely after the MCED assay data had been locked at the CRO 's end. This unblinding key enabled Geneseeq to link the de-identified assay results with the corresponding clinical data (cancer status) for analysis.

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

NCT06011694

Study protocol

The study protocol and statistical analysis plan has been submitted as a supplemental file along with the manuscript.

Data collection	Data collection details are summarized in the method section and can be found in the study protocal
Outcomes	Endpoints can be found in the protocal.

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA