

9-27-2024


## Noninvasive Multi-Cancer Detection Using Blood-Based Cell-Free MicroRNAs

Jason Zhang

Hallgeir Rui

Hai Hu

Follow this and additional works at: <https://jdc.jefferson.edu/ppcbfp>

 Part of the [Diagnosis Commons](#), [Health Services Research Commons](#), [Neoplasms Commons](#), and the [Nucleic Acids, Nucleotides, and Nucleosides Commons](#)

**[Let us know how access to this document benefits you](#)**

---

This Article is brought to you for free and open access by the Jefferson Digital Commons. The Jefferson Digital Commons is a service of Thomas Jefferson University's [Center for Teaching and Learning \(CTL\)](#). The Commons is a showcase for Jefferson books and journals, peer-reviewed scholarly publications, unique historical collections from the University archives, and teaching tools. The Jefferson Digital Commons allows researchers and interested readers anywhere in the world to learn about and keep up to date with Jefferson scholarship. This article has been accepted for inclusion in Department of Pharmacology, Physiology, and Cancer Biology Faculty Papers by an authorized administrator of the Jefferson Digital Commons. For more information, please contact: [JeffersonDigitalCommons@jefferson.edu](mailto:JeffersonDigitalCommons@jefferson.edu).



# OPEN Noninvasive multi-cancer detection using blood-based cell-free microRNAs

Jason Zhang<sup>1</sup>, Hallgeir Rui<sup>2</sup> & Hai Hu<sup>3,4</sup>✉

Patients diagnosed with early-stage cancers have a substantially higher chance of survival than those with late-stage diseases. However, the option for early cancer screening is limited, with most cancer types lacking an effective screening tool. Here we report a miRNA-based blood test for multi-cancer early detection based on examination of serum microRNA microarray data from cancer patients and controls. First, a large multi-cancer training set that included 1,408 patients across 7 cancer types and 1,408 age- and gender-matched non-cancer controls was used to develop a 4-microRNA diagnostic model using 10-fold cross-validation. In three independent validation sets comprising a total of 4,875 cancer patients across 13 cancer types and 3,722 non-cancer participants, the 4-microRNA model achieved greater than 90% sensitivity for 9 cancer types (lung, biliary tract, bladder, colorectal, esophageal, gastric, glioma, pancreatic, and prostate cancers) and 75–84% sensitivity for 3 cancer types (sarcoma, liver, and ovarian cancer), while maintaining greater than 99% specificity. The sensitivity remained to be > 99% for patients with stage 1 lung cancer. Our study provided novel evidence to support the development of an inexpensive and accurate miRNA-based blood test for multi-cancer early detection.

**Keywords** Multi-cancer early detection, MicroRNA, Noninvasive, Blood-based diagnostic model

Cancer ranks the first or second leading cause of death in most countries worldwide<sup>1</sup>. In the United States, the American Cancer Society estimated 1.9 million new cancer cases and nearly 610 K cancer deaths in 2022<sup>2</sup>. Patients diagnosed with early-stage cancers have much higher survival rates than those at late stages. For example, the 5-year patient survival rate for localized colorectal cancers is 91% but only 15% for those that have spread to distant organs<sup>2</sup>. However, early-stage cancer patients often have no symptoms and thus are more likely to miss timely diagnosis<sup>3,4</sup>. Therefore, detecting cancers at early stages is paramount to reduce cancer-related mortality.

The most effective way for detecting cancer early is the availability and accessibility of cancer screening tools for the general population. Unfortunately, the options of such screening tools are limited. Currently, only four cancer types have screening tests recommended by the United States Preventive Service Task Force (USPSTF): mammography for breast cancer, cytology/HPV testing for cervical cancer, colonoscopy and/or stool-based testing for colon cancer, and low-dose CT scans for lung cancer<sup>5–8</sup>. A challenge of using these single cancer-based screening tests is that when used sequentially, they could lead to dramatically increased cumulative incidence of false positives<sup>9</sup>. Therefore, a low cost, high performance and noninvasive test that can detect multiple cancers simultaneously will overcome the pitfalls of these single cancer-based screening tools and greatly facilitate the adoption and increase the compliance of the so-called multi-cancer early detection (MCED) in high-risk general population.

Here we report the development of a circulating microRNA (miRNA)-based MCED model using a multi-cancer training set and show its validation in a broader cohort of patients and controls, demonstrating a high accuracy of detecting 12 cancer types.

## Results

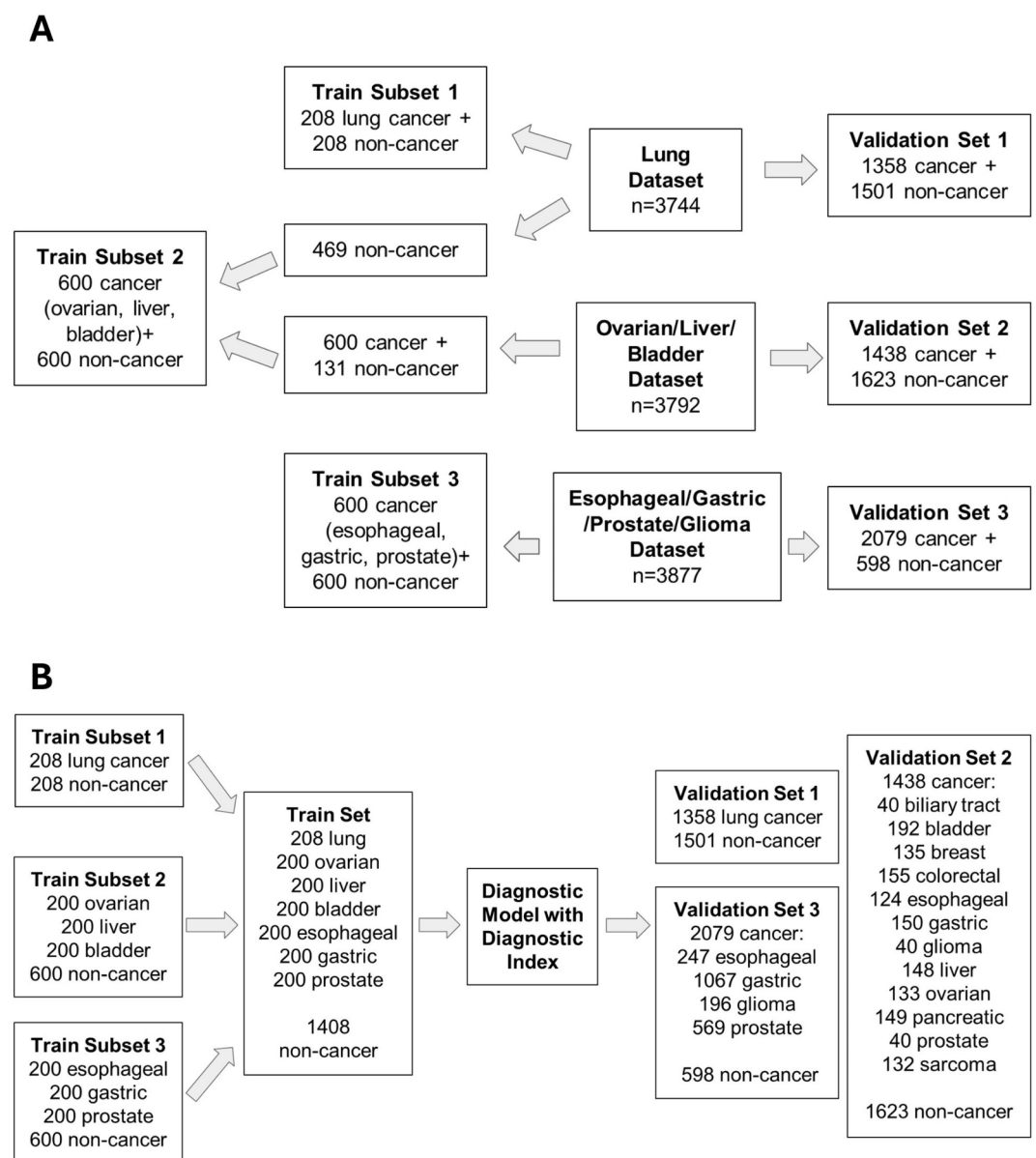
### Participants and datasets

To develop a MCED model, we identified eight serum miRNA microarray datasets from Gene Expression Omnibus (GEO)<sup>10,11</sup>, which included data from 13 cancer types (biliary tract, bladder, breast, colorectal, esophageal, gastric, glioma, liver, lung, ovarian, pancreatic, prostate, sarcoma) and were all generated from the

<sup>1</sup>Del Norte High School, San Diego, CA, USA. <sup>2</sup>Department of Pharmacology, Physiology & Cancer Biology, Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA, USA. <sup>3</sup>Chan Soon-Shiong Institute of Molecular Medicine at Windber, 620 7th Street, 15963 Windber, PA, USA. <sup>4</sup>miRoncol Diagnostics, Inc, Philadelphia, PA, USA. ✉email: h.hu@wriwindber.org

Japanese nationwide multi-year, multi-center research program “Development and Diagnostic Technology for Detection of miRNA in Body Fluids” using a standardized microarray platform. These eight datasets were originally used to develop individual diagnostic models for individual cancer types<sup>12–19</sup>. In this study, we cleaned and assembled these datasets to build a multi-cancer train set comprised 1408 cancer patients from 7 cancer types (lung, ovarian, liver, bladder, esophageal, gastric, and prostate) and 1408 age- and gender-matched non-cancer controls for the development of a diagnostic model for simultaneously detecting multiple cancer types. All the remaining subjects including 4875 cancer patients across 13 cancer types and 3722 non-cancer controls constitute three validation sets. Detailed description on study design, microarray datasets and construction of train and validation datasets are described in the Supplemental Methods and Fig. 1.

Detailed demographic and clinical information for those cancer types of large sample size were described in the original publications. Briefly, the patients in the lung cancer dataset ( $n=1566$ ) had mean age 65y, composed of 57% male and 62% former or current smokers, with 78% of the tumors being adenocarcinoma, 14% squamous carcinoma, 87% stage I or II. The bladder cancer dataset ( $n=392$ ) included patients of mean age 68y, 72% male, 95% non-metastatic, 88% nodal-negative, 77% T1 and 80% high grade. The ovarian cancer dataset ( $n=333$ ) included patients with mean age 57y, 35% stage I or II, 96% epithelial (including 55%, 19% and 13% for serous, clear cell, and endometrioid histology, respectively). The patients in the liver cancer dataset ( $n=348$ ) were of mean age 68y, 78% male, and 70% stage I or II. The esophageal cancer dataset ( $n=447$ ) consisted of patients with a mean age 67y, 97% male, and 66% stage I or II. The gastric cancer dataset ( $n=1267$ ) included patients



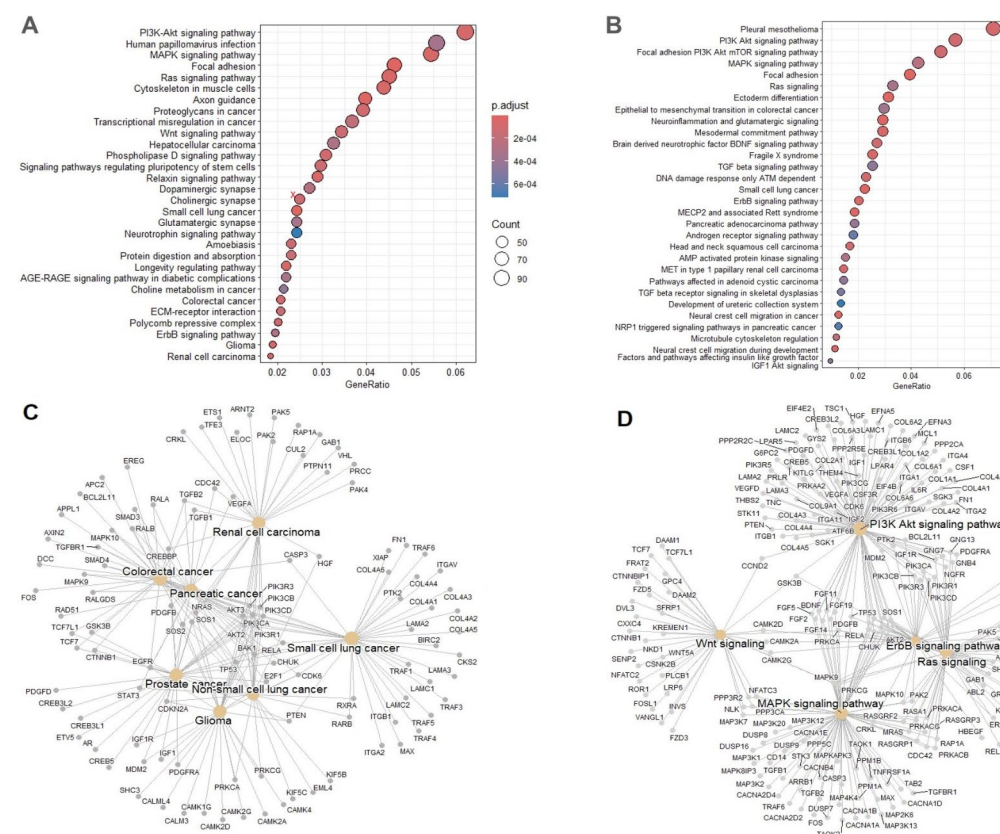
**Fig. 1.** Flow of datasets and study design. **(A)** Construction of the train and validation datasets. **(B)** Study design of model development and validation.

with mean age 66y, 77% male and all stage I or II. The glioma dataset ( $n = 196$ ) comprised patients with mean age 56y and 57% were male. Finally, the patients in the prostate cancer dataset ( $n = 769$ ) had mean age 68y, 93% node-negative, and 92% non-metastatic.

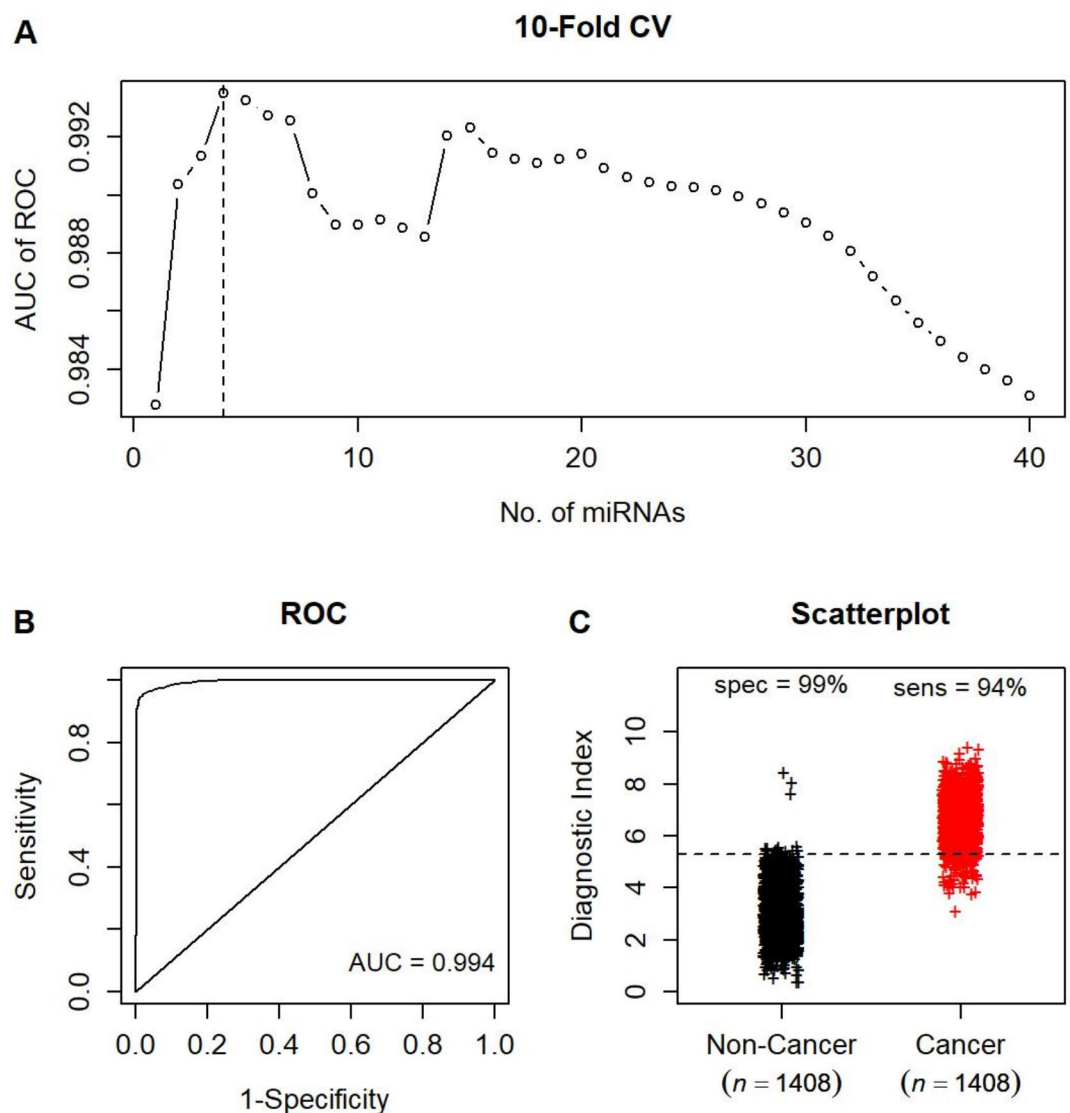
## Cancer diagnostic model development

All diagnostic model development work was performed in the multi-cancer Train Set, which included 1408 cancer patients and 1408 non-cancer controls matched by age and gender (Fig. 1B). First, limma analysis was used to assess the differential expression of miRNAs between cancer and non-cancer. miRNAs were then ranked based on the B statistics from the limma analysis. The top 50 differentially expressed miRNAs are listed in Supplemental Table 1. A key hallmark of cancer is uncontrolled proliferation. We hypothesized that cancer-associated miRNAs would target growth signaling pathways. To investigate this, potential targeted genes regulated by these top 50 differentially expressed miRNAs were predicted from the miRDB database. Both KEGG and WikiPathways analysis of potential target genes indicated that several cancer pathways including colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, renal cell carcinoma, glioma etc. were enriched (Fig. 2A, B and C). Consistent with our hypothesis, known signal transduction pathways implicated in tumorigenesis and cancer progression including PI3K Akt signaling, MAPK signaling, Wnt signaling, ErbB signaling, Ras signaling, etc. were enriched as well (Fig. 2A, B and D). Furthermore, network analysis of enriched pathways showed that common target genes were involved among several cancer and signaling pathways (Fig. 2C and D), which supports the implication of common miRNAs for the diagnosis of multiple cancer types.

Ten-fold cross-validation revealed that the top 4 miRNAs (hsa-miR-5100, hsa-miR-1228-5p, hsa-miR-8073 and hsa-miR-663a) provided the highest AUC in the ROC analysis and thus were included in the final diagnostic model (Fig. 3A). We calculated a diagnostic index by the weighted sum of the 4 miRNA expression levels (weighted by the t statistics from the limma analysis) and normalized to the range of 0 to 10. This 4-miRNA model achieved an AUC value of 0.994 within the Train set (Fig. 3B). A cut-point of 5.3 was chosen to yield an overall > 99% specificity (i.e., < 1% false positives) across the non-cancer cases, and an overall 94% sensitivity (Fig. 3C). The AUC and sensitivity of the model for each of the 7 cancer types in the multi-cancer Train set ranged from 0.985 to 84% for ovarian cancer to 0.998 and 100% for bladder and gastric cancers, respectively (Table 1).



**Fig. 2.** Pathway enrichment analysis of target genes regulated by the top 50 differentially expressed miRNAs. (A) KEGG analysis of potential target genes regulated by these miRNAs<sup>44,45</sup>. (B) WikiPathways analysis of potential target genes regulated by these miRNAs<sup>46</sup>. (C) Network plot of enriched target genes depicting the linkages of genes and selected KEGG pathways<sup>44,45</sup>. (D) Network plot of enriched target genes depicting the linkages of genes and selected WikiPathways<sup>46</sup>. For (A) and (B), the gene ratio (no. of mapped genes / total no. of genes) is shown on the x-axis; bubble size is gene count and bubble color reflects adjusted p value.



**Fig. 3.** Diagnostic performance of the 4-miRNA model in the multi-cancer Train Set. **(A)** 10-fold cross validation; **(B)** ROC of the 4-miRNA model; **(C)** Scatterplot of the diagnostic index.

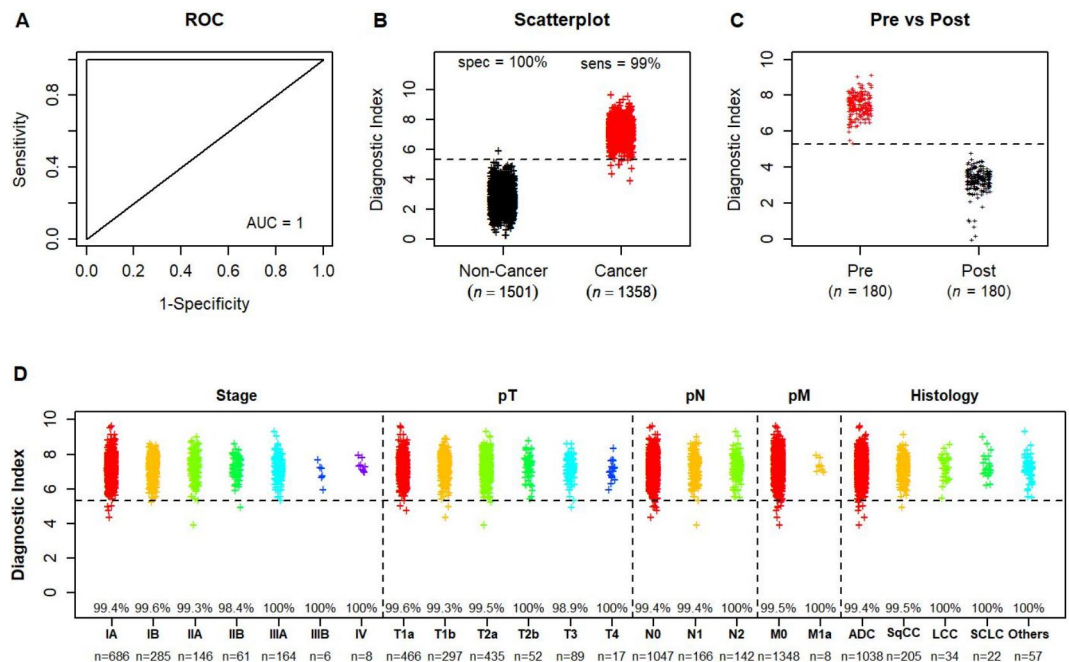
Cancer types	N	AUC of ROC	Sensitivity
Lung	208	0.997	99%
Bladder	200	0.998	100%
Ovarian	200	0.985	84%
Liver	200	0.987	86%
Gastric	200	0.998	100%
Esophageal	200	0.993	92%
Prostate	200	0.997	97%

**Table 1.** Performance of the 4-miRNA model for each cancer type in the multi-cancer train set.

### Validation of the diagnostic model in the independent validation set 1

The performance of the 4-miRNA model was first evaluated in the independent Validation Set 1 ( $n = 2859$ ) that included 1358 lung cancer patients and 1501 non-cancer controls. The model achieved an AUC of 1.000 (Fig. 4A) with a specificity of 100% and sensitivity of 99% (Fig. 4B). In addition, analysis of paired serum samples (pre- vs. post-surgery;  $n = 180$ ) verified normalization of the diagnostic indices to the levels of non-cancer controls in post-surgery serum samples (Fig. 4C).





**Fig. 4.** Diagnostic performance of the 4-miRNA model in Validation Set 1, the lung cancer validation dataset. **(A)** ROC of the 4-miRNA model; **(B)** Scatterplot of the diagnostic index; **(C)** Scatterplot of the diagnostic index from pre- vs. post-operation serum samples; **(D)** Scatterplot of the diagnostic index in clinical subsets. ADC: adenocarcinoma; SqCC: squamous cell carcinoma; LCC: large cell carcinoma; SCLC: small cell lung cancer.

Furthermore, the performance of the 4-miRNA model was evaluated across clinical subsets of the Validation Set 1, as defined by the clinical stages, TNM stages, and histology subtypes. High sensitivities were observed for all clinical subsets. The model achieved at least 99% sensitivity for 22 out of 24 clinical subsets examined except for stage IIB and T3 tumors (Fig. 4D). In particular, the model demonstrated > 99% sensitivities for stage I lung cancers and for adenocarcinoma and squamous cell carcinoma.

### Validation of the diagnostic model in the independent validation sets 2 and 3

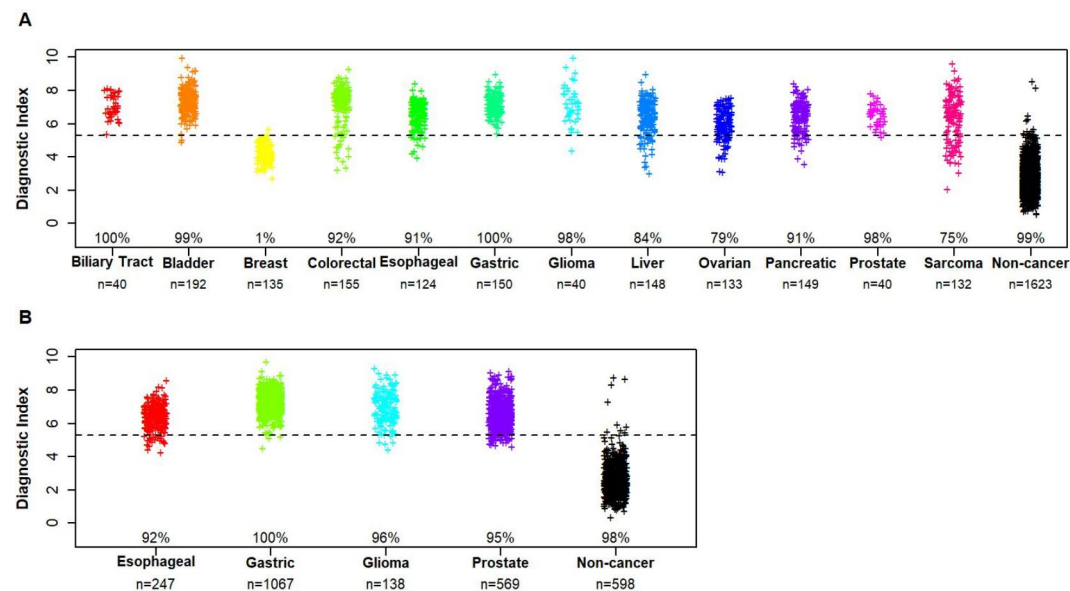
The independent Validation Set 2 included 1438 patients across 12 additional cancer types and 1623 non-cancer controls. Except for breast cancer, the 4-miRNA model achieved at least 90% sensitivity for eight cancer types (biliary tract, bladder, colorectal, esophageal, gastric, glioma, pancreatic and prostate) and at least 75% for the other three cancer types (liver, ovarian and sarcoma) (Fig. 5A; Table 2). Noteworthy, while the model had a reasonable AUC value of 0.909 for breast cancer, the 1% sensitivity was still very low due to the high specificity requirement (Fig. 5A; Table 2).

The independent Validation Set 3 included 2079 patients from four cancer types (esophageal, gastric, glioma and prostate) and 598 non-cancer controls, where the sample sizes of the four cancer types were substantially larger than those in Validation Set 2 (247 vs. 124 for esophageal, 1067 vs. 150 for gastric, 196 vs. 40 for glioma, and 569 vs. 40 for prostate). The 4-miRNA model achieved > 0.99 AUC and > 99% sensitivity for all four cancer types, similar to those observed in Validation Set 2 (Fig. 5B; Table 2). The specificity of the model was a little lower in Validation Set 3 than in Validation Set 2 (0.98 vs. 0.99) (Fig. 5B). Therefore, for Validation Set 3, a sensitivity analysis with an adjusted diagnostic index cut-point of 5.6 was explored to increase the specificity of the new model to 99%. With this new cut-point, the model still achieved high sensitivity for all four cancer types, including 99% for gastric, 92% for glioma, 91% for prostate, and 89% for esophageal cancers.

### Discussion

Noninvasive screening tests for MCED via analyzing circulating cell-free nucleic acids and/or proteins in the body fluid, especially blood, have attracted high attention for the last decade. In this study, we reported the development and validation of a serum 4-miRNA diagnostic model and demonstrated that in three large independent validation sets totaling 8597 participants (4875 cancer patients across 13 cancer types and 3722 non-cancer individuals), the 4-miRNA model can detect 12 cancer types simultaneously with high sensitivities (> 90% for 9 cancer types, and  $\geq 75\%$  for 3 cancer types) while still achieving a very high specificity of  $\sim 99\%$ . In addition, the observation that the diagnostic indices for the post-surgery serum samples were reduced to normal levels suggests the potential utility of the model for monitoring response to treatment and detection of recurrence.

Importantly, our model was able to detect early-stage cancers at high sensitivity. Specifically, in Validation Set 1 of lung cancer patients, the model detects stage I and II cancers at a sensitivity ranging from 98.4 to 99.6%



**Fig. 5.** Diagnostic performance of the 4-miRNA model in Validation Sets 2 and 3. **(A)** Scatterplot of the diagnostic index in Validation Set 2; **(B)** Scatterplot of the diagnostic index in Validation Set 3.

	N	AUC of ROC	Sensitivity
Validation Set 2			
Biliary tract	40	0.998	100%
Bladder	192	0.999	99%
Breast	135	0.909	1%
Colorectal	155	0.991	92%
Esophageal	124	0.996	91%
Gastric	150	0.999	100%
Glioma	40	0.997	98%
Liver	148	0.998	84%
Ovarian	133	0.986	79%
Pancreatic	149	0.995	91%
Prostate	40	0.998	98%
Sarcoma	132	0.976	75%
Validation Set 3			
Gastric	1067	0.994	100%
Glioma	196	0.993	96%
Esophageal	247	0.992	92%
Prostate	569	0.993	95%

**Table 2.** Performance of the 4-miRNA model for each cancer type in Validation Sets 2 and 3.

(Fig. 4D). In Validation Sets 2 and 3, while individual patient-level stage information was not available, aggregate stage information was provided for 6 of the 12 cancer types examined. First, all gastric cancer patients were stage I or II, thus the 100% sensitivity of our model applied to early-stage gastric cancer. Second, 88% and 93% of bladder and prostate cancer patients had node negative disease. Thus, with 99% and 98% sensitivity for these two cancers, the sensitivity for stage I or II bladder and prostate cancers should be very high as well. Third, 66% and 70% of esophageal and liver cancer patients were stage I or II, respectively. It was reasonable to speculate that the sensitivity for stage I or II of these two cancers should not be far off from the 92% and 84% sensitivity reported for all stages included. In summary, based on the data currently available in the three validation sets, we concluded that our 4-miRNA model achieves high sensitivity for stage I or II disease of six cancer types (lung, gastric, bladder, prostate, esophageal, and liver).

Of note, the original studies that generated the eight miRNA microarray datasets analyzed in this study also proposed miRNA panels for detecting each of the eight cancer types (lung, ovarian, liver, bladder, esophageal squamous, gastric, prostate and glioma), respectively. These eight miRNA panels included 41 unique miRNAs with only one overlapping miRNA, hsa-miR-6724-5p, which occurred in the liver and bladder cancer panels.

While some of these panels demonstrated higher performance characteristics than our model for their respective cancer types, this is expected given their specific focus. However, if these panels were to be used to detect these eight cancer types together in a sequential fashion, the cumulative incidence of false positives was approximately 33% based on the published performance matrix. In contrast, our model, which detects 12 cancer types simultaneously, achieves a false positive rate of less than 1%.

Among the four miRNAs used in our model, hsa-miR-5100 has been reported to be overexpressed in lung, gastric, oral squamous cell carcinoma, and pancreatic cancers<sup>20–24</sup>. On the other hand, hsa-miR-1228-5p has been implicated as overexpressed in hepatocellular carcinoma and kidney clear cell carcinoma<sup>25,26</sup>, while hsa-miR-663a has been found to be overexpressed in colon cancer and metastatic prostate cancer<sup>27,28</sup>. Gene set enrichment and network analysis showed that transforming growth factor beta-1 (TGFB1), a gene regulated by hsa-miR-663a, was implicated in signaling pathways across multiple cancer types including colorectal cancer, pancreatic cancer, gastric cancer, renal cell carcinoma, hepatocellular carcinoma and leukemia. The observation that the PI3K Akt and MAPK signaling pathways are among the most regulated by the top 50 miRNAs certainly suggests that the origin of the miRNAs is from the cancer cells, but not from reactive stromal fibroblasts, tumor-associated immune cells, or biopsy-induced wound-related changes<sup>29</sup>. Taken together, these data support the use of these miRNAs as potential biomarkers for cancer early detection across multiple cancer types.

Several commercial assays for MCED have emerged in recent years. Most of these tests used next generation sequencing (NGS) technology to evaluate either methylation or fragmentation patterns of circulating tumor DNAs<sup>30–33</sup>. The most prominent MCED test that attracted high attention was the Galleri test that examined > 100,000 targeted methylated regions and > 1,000,000 CpG dinucleotides. In its prospective and case-controlled the Circulating Cell-free Genome Atlas (CCGA) study, Galleri achieved an overall sensitivity of 67.6% across 12 stage I–III pre-specified cancer types and 99.5% specificity<sup>30</sup>. However, the sensitivity was only 16.8% for stage I and 40.4% for stage II. The other MCED test not based on NGS technology is CancerSEEK that assesses four biomarker classes (aneuploidy, DNA methylation, mutations and proteins). In its latest retrospective, case-controlled study of 566 cancer patients across 12 cancer types and 566 non-cancer controls, it showed an overall 61% sensitivity and 98.2% specificity<sup>34</sup>. The sensitivity dropped to 49.8% for stage I–III cancers. In summary, these MCED tests generally showed modest sensitivities in the range of 60–70% when a high 99% specificity was required, and the sensitivities dropped further for stage I or II cancers. Compared to these assays, our diagnostic model, while much simpler, demonstrated substantially higher sensitivities in the range of 90–100% for 9 out of 12 cancer types in large validation cohorts totaling almost 8600 participants. More importantly, our model achieves similarly high sensitivities for stage I or II cancers.

The clinical utility of these MCED assays must be ultimately demonstrated in prospective screening trials with asymptomatic individuals. For example, Galleri was evaluated in the prospective screening study of PATHFINDER that analyzed 6621 participants aged  $\geq 50$  years with 1 year follow-up<sup>35</sup>. The study detected a cancer signal in 92 (1.4%) participants and confirmed 35 as true positives, resulting in a 38% positive predictive value (PPV). In addition, 121 participants had cancer diagnosed at the end of 1-year follow-up, which corresponded to a 29% sensitivity by Galleri. In its latest prospective observation study SYMPLIFY with 5461 symptomatic participants referred from primary care and 368 (6.7%) diagnosed with a cancer, Galleri achieved 66.3% sensitivity and 98.4% specificity<sup>36</sup>. For our 4-miRNA diagnostic model, assuming a screening population with 1% cancer incidence rate, 90% sensitivity and 99.3% specificity, our model would provide a PPV of 56%, significantly higher than the 3.7–4.4% PPVs for the four single-cancer screening tests recommended by USPSTF<sup>37–39</sup>.

It is worth noting that a simple four-parameter diagnostic model like the one described here not only costs significantly less, but also can be developed into an in vitro diagnostic (IVD) test using RT-qPCR capable of decentralized testing, which has an advantage over NGS-based tests that are usually implemented as a laboratory developed test (LDT). These characteristics are important to drive adoption and increase affordability of MCED tests as they are intended to target high risk or at-risk general public, especially for those from low-income communities.

We acknowledge that the current study is a computational analysis using public datasets. Experimental validation and investigations on the role of the 4 miRNAs will shed light on the mechanistic understanding of the predictive power of these miRNAs. In particular, validating these miRNAs in different cohorts using different molecular techniques such as PCR is crucial before considering the current study results definitive, which is also critically important in developing these miRNAs into a lower-cost and practical diagnostic assay for clinical use. These will be the focus of our future work and are beyond the scope of the current study.

In summary, our study has provided proof-of-concept data for developing a blood screening test based on expression profiles of circulating cell-free miRNAs for 12 cancer types, which account for 50% estimated new cancer cases and 63% cancer deaths in the US in 2022<sup>2</sup>.

## Methods

### *Study design and construction of train and validation datasets*

We identified eight serum miRNA microarray datasets from Gene Expression Omnibus (GEO)<sup>10,11</sup>. After removing redundant cases, we assembled three large datasets that were independent of each other: a lung cancer dataset ( $n = 3744$ )<sup>10,12</sup>, a combined dataset by merging the ovarian, liver and bladder cancer datasets ( $n = 3792$ )<sup>10,13–15</sup>, and a combined dataset by merging the esophageal squamous cell, gastric, prostate and glioma cancer datasets ( $n = 3877$ )<sup>11,16–19</sup>.

Based on these three large datasets, we constructed a large training set ('Train Set') that included 1408 cancer patients from 7 cancer types (208 lung cancer patients and 200 patients each for ovarian, liver, bladder, esophageal, gastric, and prostate) and 1408 age- and gender-matched non-cancer controls for the development of a diagnostic model for detecting multiple cancer types. All the remaining cases formed three separate



independent validation sets (Fig. 1A and B). Details of how the cancer case and control samples for the Train Set and Validation Sets were selected are described in the Supplemental Methods.

### Blood sample collection

Collection of blood serum samples has been previously described in the original publications<sup>12–19</sup>. Briefly, serum samples were collected prior to surgical operation from cancer patients who were admitted to the National Cancer Center Hospital (NCCH) between 2008 and 2016 and stored initially at 4°C for one week and then at -20°C until further use. The exclusion criteria included those patients who were treated with preoperative chemotherapy and/or radiotherapy prior to serum sample collection. The serum samples for non-cancer controls were from those who had no history of cancer and no hospitalization during the previous 3 months and were collected along with routine blood tests from outpatient departments of three sources: NCCH, National Center for Geriatrics and Gerontology (NCGG) Biobank and Yokohama Minoru Clinic (YMC). Serums from cancer patients and non-cancer controls collected at NCCH were stored in the same way as described above, while those from NCGG and YMC were stored at -80°C till use. While our study as an *in silico* analysis of public datasets does not require any ethical approval, the original studies were approved by the NCCH Institutional Review Board, the Ethics and Conflict of Interest Committee of the NCGG, and the Research Ethics Committee of Medical Corporation Shintokai YMC. Written informed consent was obtained from each participant<sup>12–19</sup>.

### Microarray analysis of miRNA expression

Details about microarray expression analysis were described in the original publications<sup>12–19</sup>. Briefly, total RNA was extracted from 300 µl serum, labeled by 3D-Gene miRNA Labeling kit and hybridized to 3D-Gene Human miRNA Oligo Chip (Toray Industries, Kanagawa, Japan) that evaluates the expression profiles of 2588 miRNA sequences registered in miRBase release 21 (<http://www.mirbase.org/>). Low quality samples were discarded if the coefficient of variation of negative control probes > 0.15 or the number of flagged probes identified by 3D-Gene Scanner as “uneven spot images” > 10. A miRNA was called “present” when its signal intensity was greater than mean plus two standard deviations of the negative control signals after the top and bottom 5% of the ranked signal intensities were removed. The signal intensities for miRNAs were determined after background subtraction by subtracting the mean signal intensity of negative control signals (after removing top and bottom 5% of the ranked signal intensities) from the miRNA signal. Finally, microarrays were normalized by calibrating according to three pre-selected internal control miRNAs (miR-149-3p, miR-2861, and miR-4463).

### Diagnostic model development

miRNA biomarker identification and all model development work were done in the multi-cancer Train Set only. The differential miRNA expression between cancer vs. non-cancer was evaluated using Linear Model for Microarray Data (limma)<sup>40</sup>. miRNAs were then ranked based on the *t* statistics from the limma analysis and the top miRNAs were used to build diagnostic models for distinguishing cancer vs. non-cancer. A diagnostic index was calculated for each diagnostic model as a linear sum of expression levels of the selected miRNAs weighted by limma statistics. Ten-fold cross validation was performed to determine the optimal number of miRNAs to be included in the final diagnostic model that had the highest area-under-the-curve (AUC) of the Receiver Operating Characteristics (ROC) curves for distinguishing cancer vs. non-cancer. The cut-point for the diagnostic index was chosen to ensure at least 99% specificity (i.e., ≤ 1% false positive rate) as the model may potentially be used as a screening tool in at-risk general public.

### Diagnostic model validation

The three independent validation datasets contained mutually exclusive samples that were not used in model development, with each offering distinct characteristics for the validation of the developed model. Validation Set 1 not only was of a very large sample size for lung cancer cases, but also contained comprehensive patient-level clinicopathologic data in contrast to the other two validation datasets, making it possible to assess model performance on early-stage cancers and different histology subtypes. Validation Set 2 contained samples from 12 other cancer types, thus expanding the evaluation of model performance across multiple cancer types. Validation Set 3 comprised large numbers of the cases from four cancer types including the two cancer types with low sample size in Validation Set 2, allowing additional independent verification of the model performance.

### KEGG and WikiPathways pathway enrichment analysis of potential miRNAs target genes

The prediction of target genes of top 50 significantly differentially expressed miRNAs was performed using the database miRDB<sup>41</sup>. The pathway enrichment analysis on the target genes was conducted using Bioconductor package clusterProfiler (version 4.10.1)<sup>42,43</sup> based on KEGG<sup>44,45</sup> and WikiPathways<sup>46</sup>. A Benjamini–Hochberg adjusted *p* value cutoff 0.05 was used to select significantly enriched pathways.

### Statistical analysis

AUC of the ROC curve analysis, sensitivity, and specificity were used to measure the diagnostic performance for detecting cancer vs. non-cancer. Sensitivity was defined as the proportion of cancer patients who were correctly identified as cancer by the diagnostic model, while specificity was defined as the proportion of non-cancer participants who were correctly identified as non-cancer. limma analysis was performed using Bioconductor package limma (<http://www.bioconductor.org/>)<sup>40</sup>. All statistical analysis was conducted using R version 4.2.1 (<http://www.r-project.org/>).

## Data availability

All individual patient data were made publicly available by the original study authors. Gene Expression Omnibus (GEO) accession IDs for the datasets used in this study are included in the Supplemental Methods section.

Received: 17 June 2024; Accepted: 20 September 2024

Published online: 27 September 2024

## References

- Sung, H. et al. Global Cancer Statistics: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**(3), 209–249 (2021).
- Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**(1), 7–33 (2022).
- Ahlquist, D. A. Universal cancer screening: revolutionary, rational, and realizable. *NPJ Precis Oncol.* **2**, 23 (2018).
- Noone, A. M. et al. (eds) SEER Cancer Statistics Review, 1975–2015, National Cancer Institute. (2018).
- Siu, A. L. & U.S. Preventive Services Task Force. Screening for breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **164** (4), 279–296 (2016).
- US Preventive Services Task Force et al. Screening for cervical Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA.* **320**(7), 674–686 (2018).
- US Preventive Services Task Force et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA.* **325**(19), 1965–1977 (2021).
- US Preventive Services Task Force et al. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA.* **325**(10), 962–970 (2021).
- Croswell, J. M. et al. Cumulative incidence of false-positive results in repeated, multimodal cancer screening. *Ann. Fam. Med.* **7**(3), 212–222 (2009).
- Zhang, A. & Hu, H. A novel blood-based microRNA diagnostic model with high accuracy for Multi-cancer Early Detection. *Cancers (Basel).* **14**(6), 1450 (2022).
- Zhang, A. & Hu, H. Independent validation of a novel noninvasive 4-microRNA diagnostic model for multicancer early detection. *J. Clin. Oncol.* **40**(16\_suppl), 3065–3065 (2022).
- Asakura, K. et al. A miRNA-based diagnostic model predicts resectable lung cancer in humans with high accuracy. *Commun. Biol.* **3**(1), 134 (2020).
- Yokoi, A. et al. Integrated extracellular microRNA profiling for ovarian cancer screening. *Nat. Commun.* **9**(1), 4319 (2018).
- Yamamoto, Y. et al. Highly sensitive circulating MicroRNA Panel for Accurate Detection of Hepatocellular Carcinoma in patients with Liver Disease. *Hepatol. Commun.* **4**(2), 284–297 (2020).
- Usuba, W. et al. Circulating miRNA panels for specific and early detection in bladder cancer. *Cancer Sci.* **110**(1), 408–419 (2019).
- Sudo, K. et al. Development and validation of an esophageal squamous cell carcinoma detection model by large-scale MicroRNA profiling. *JAMA Netw. Open.* **2**(5), e194573 (2019).
- Abe, S. et al. A novel combination of serum microRNAs for the detection of early gastric cancer. *Gastric Cancer.* **24**(4), 835–843 (2021).
- Ohno, M. et al. Assessment of the diagnostic utility of serum MicroRNA classification in patients with diffuse glioma. *JAMA Netw. Open.* **2**(12), e1916953 (2019).
- Urabe, F. et al. Large-scale circulating microRNA profiling for the liquid biopsy of prostate Cancer. *Clin. Cancer Res.* **25**(10), 3016–3025 (2019).
- Huang, H. et al. miR-5100 promotes tumor growth in lung cancer by targeting Rab6. *Cancer Lett.* **362**(1), 15–24 (2015).
- Wang, T., Liu, X., Tian, Q., Liang, T. & Chang, P. Increasing expression of miR-5100 in non-small-cell lung cancer and correlation with prognosis. *Eur. Rev. Med. Pharmacol. Sci.* **21** (16), 3592–2597 (2017).
- Zhang, H. M. et al. MKL1/miR-5100/CAAP1 loop regulates autophagy and apoptosis in gastric cancer cells. *Neoplasia.* **22**(5), 220–230 (2020).
- Chijiwa, Y. et al. Overexpression of microRNA-5100 decreases the aggressive phenotype of pancreatic cancer cells by targeting PODXL. *Int. J. Oncol.* **48**(4), 1688–1700 (2016).
- Wei, Z., Lyu, B., Hou, D. & Liu, X. Mir-5100 mediates Proliferation, Migration and Invasion of oral squamous cell carcinoma cells Via Targeting SCAI. *J. Invest. Surg.* **34**(8), 834–841 (2021).
- Tan, Y. et al. A serum MicroRNA panel as potential biomarkers for Hepatocellular Carcinoma related with Hepatitis B Virus. *PLoS One.* **9**(9), e107986 (2014).
- Shen, J. et al. Comprehensive analysis of expression profiles and prognosis of TRIM genes in human kidney clear cell carcinoma. *Aging.* **14**(10), 4606–4617 (2022).
- Qin, D. et al. A circulating miRNA-Based Scoring System established by WGCNA to predict Colon cancer. *Anal. Cell. Pathol.* **2019**, 1–7 (2019).
- Knyazev, E. N. et al. Shkurnikov MYu. MicroRNA hsa-miR-4674 in hemolysis-free blood plasma is Associated with distant metastases of Prostatic Cancer. *Bull. Exp. Biol. Med.* **161**(1), 112–115 (2016).
- Kameyama, H. et al. Needle biopsy accelerates pro-metastatic changes and systemic dissemination in breast cancer: implications for mortality by surgery delay. *Cell. Rep. Med.* **4**(12), 101330 (2023).
- 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**(9), 1167–1177 (2021).
- Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science.* **359**(6378), 926–930 (2018).
- Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **11**(1), 3475 (2020).
- Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature.* **570**(7761), 385–389 (2019).
- Douville, C. et al. Multi-cancer early detection through evaluation of aneuploidy, methylation, and protein biomarkers in plasma. *Ann. Oncol.* **33**(S\_7), S575 (2022).
- Schrag, D. et al. A prospective study of a multi-cancer early detection blood test. *Ann. Oncol.* **33**(S\_7), S961 (2022).
- Nicholson, B. D. et al. Multi-cancer early detection test in symptomatic patients referred for cancer investigation in England and Wales (SYMPHONY): a large-scale, observational cohort study. *Lancet Oncol.* **24**(7), 733–743 (2023).
- Lehman, C. D. et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the breast Cancer Surveillance Consortium. *Radiology.* **283**(1), 49–58 (2017).
- U. S. Food and Drug Administration. Cologuard Summary of Safety and Effectiveness Data (Premarket Approval Application P130017). (2014).
- National Lung Screening Trial Research Team et al. Results of initial low-dose computed tomographic screening for lung cancer. *N Engl. J. Med.* **368**(21), 1980–1991 (2013).
- Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**(7), e47 (2015).

41. Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* **48**(D1), D127–D131 (2020).
42. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* **16**(5), 284–287 (2012).
43. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov. (Cambridge (Mass)).* **2**(3), 100141 (2021).
44. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
45. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**(D1), D587–D592 (2023).
46. Agrawal, A. et al. WikiPathways 2024: next generation pathway database. *Nucleic Acids Res.* **52**(D1), D679–D689 (2024).

### Author contributions

J.Z. and H.H. conceived and designed the study. J.Z. collected and analyzed the data. J.Z., H.R., and H.H. interpreted the data and wrote and finalized the manuscript.

### Declarations

### Competing interests

J.Z. and H.H. are named inventors on a patent of the diagnostic model developed in this study. HH is a cofounder of, and holds equity in miRoncol Diagnostics, Inc, a company that seeks to commercialize the diagnostic model. All other authors do not hold any competing interest.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-73783-0>.

**Correspondence** and requests for materials should be addressed to H.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024