npj Precision Oncology

https://doi.org/10.1038/s41698-026-01620-w

Article in Press

Non-invasive esophageal cancer screening in high-risk populations: a multi-omics integration model

Received: 27 May 2025

Accepted: 13 July 2026

Published online: 23 July 2026

Cite this article as: Li, W., Cui, Y., Wu, J. *et al.* Non-invasive esophageal cancer screening in high-risk populations: a multi-omics integration model. *npj Precis. Onc.* (2026). https://doi.org/10.1038/s41698-026-01620-w

Wenya Li, Youbin Cui, Jingcai Wu, Yongzhi Wang, Huidong Yu, Xiaotong Xi, Yu Ding, Peipei Zhang, Lei Ye, Xiaoqiang Wang, Xiaohuan Hou, Zhenghua Liu, Ling Lv, Wanfu Men, Xiangli Liu, Yang Liu, Mingrui Shao, Xing Zhang, Wanglong Deng, Dongsheng Chen, Yong Li & Shun Xu

Non-Invasive Esophageal Cancer Screening in High-Risk Populations: A Multi-Omics Integration Model

Wenya Li[1*#], Youbin Cui[2*], Jingcai Wu[3*], Yongzhi Wang[4*], Huidong Yu[1*], Xiaotong Xi[5], Yu Ding[5], Peipei Zhang[5], Lei Ye[5], Xiaoqiang Wang[5], Xiaohuan Hou[6], Zhenghua Liu[1], Ling Lv[1], Wanfu Men[1], Xiangli Liu[1], Yang Liu[1], Mingrui Shao[1], Xing Zhang[5], Wanglong Deng[5], Dongsheng Chen[5#], Yong Li[7, 8#], Shun Xu[1#]

[1] Department of Thoracic Surgery, The First Hospital of China Medical University, Shenyang, 110122, China.

[2] Department of Thoracic Surgery II, Department of Lung Transplantation, Organ Transplantation Center, the First Hospital of Jilin University, Changchun, 130021, China.

[3] Chifeng Traditional Chinese Medicine and Mongolian Medicine Hospital, Chifeng, 024000, China.

[4] Department of Thoracic Surgery, Dandong First Hospital, Dandong, 118000, China.

[5] State Key Laboratory of Neurology and Oncology Drug Development, Jiangsu Simcere Diagnostics Co., Ltd., Nanjing Simcere Medical Laboratory Science Co., Ltd., Nanjing, China.

[6] Chifeng Maternity and Gynecology Hospital, Chifeng, 024000, China.

[7] Department of Thoracic Surgery,National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital,Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021,China.

[8] Central Laboratory, Hebei Collaborative Innovation Center of Tumor Microecological Metabolism Regulation, Affiliated Hospital of Hebei University, Baoding, Hebei 071000, China.

[*] First Author: Wenya Li, Youbin Cui, Jingcai Wu, Yongzhi Wang and Huidong Yu are primary authors and contributed equally.

[#] Corresponding Author: Dongsheng Chen(dongsheng.chen@simceredx.com), Yong Li(Email), Wenya Li(41578203@qq.com) and Shun Xu(xs_ydy@126.com) are both corresponding authors, Shun Xu serving as the primary contact for correspondence.

Shun Xu is the paper's lead contact(xs_ydy@126.com)

Abstract

Esophageal cancer (ESCA) is a leading cause of cancer-related mortality in China, with over 95% of patients diagnosed at advanced stages, resulting in low survival rates. Early detection is critical for improving outcomes, yet current screening methods are invasive or lack sensitivity. We developed a non-invasive early-detection model by integrating circulating cell-free DNA (cfDNA) methylation and fragmentomics features. Using a three-layer stacked ensemble machine learning framework, we analyzed multicenter blood samples from 309 participants, including ESCA patients and high-risk individuals. The integrated model called ESim-seq (Esophageal Cancer Screening with Integrated Model)achieved an AUC of 0.994 (95% CI: 0.979–1.000) in validation. At 99% specificity, sensitivity reached 87.3%, with simulated screening showing a 24.8% increase in stage I detection. Stratified analyses confirmed robustness across demographics, risk factors, and tumor stages. Clinical benefit modeling projected a 56.3% stage I detection rate under ideal conditions, significantly improving survival outcomes compared to standard care. This study developed a cfDNA-based analytical framework and demonstrated the advantage of multi-feature integration for early detection of ESCA.

Keywords: Esophageal cancer, early detection, high-risk population, multi-omics integration, cell-free DNA

Introduction

Esophageal cancer(ESCA), a common malignancy, gravely endangers Chinese residents' health. In 2022, China saw 224,000 new ESCA cases and 187,500 deaths, accounting for 4.64% and 7.28% of all malignant tumors respectively[1]. Although its incidence shows a downward trend, early detection rate is very low[2]. ESCA often lacks distinctive clinical symptoms in the early stage, making it easy to be overlooked. Over 95% of patients are diagnosed with mid-to-late stage cancer during their first visit, resulting in a 5-year survival rate below 25%[3-4]. However, early (Tis-T1 stage) patients receiving standard treatment like endoscopic resection can have their 5-year survival rate rise to over 90%[5]. National Cancer Center data in 2022 revealed that high-risk individuals (aged >40 with any one of these: family history of ESCA, long-term smoking/drinking, residing in high-incidence areas, preference for hot or pickled foods) have an annual incidence rate of 128.6/100,000, 12.8 times that of the general population[6]. Thus, regular screening and early treatment for high-risk groups are crucial for improving early diagnosis and reducing mortality.

Screening is vital for early cancer and precancerous lesion detection in high-risk ESCA populations. Endoscopic examination, though effective for early lesion detection, has limitations such as invasiveness, patient discomfort, and resource constraints[7]. Liquid biopsy based on circulating cell-free DNA (cfDNA) offers a new approach to break through traditional screening limitations. cfDNA, released from tumor cells, carries diverse molecular information, including mutations, copy number variations, fragmentation features, and epigenetic modifications[8]. Compared to conventional methods, it's less invasive and easier for sample collection. ctDNA mutation detection, through ultra-high-depth sequencing (>30,000×), can identify tumor signals with 0.1% variant allele frequency (VAF), yet its clinical use is hindered by clonal hematopoiesis[9] and insufficient specificity. Fragmentomics, using low-depth whole-genome sequencing (LP-WGS), captures cfDNA fragmentation features. The fragmentation patterns in cancer patients exhibits characteristically shorter fragment lengths compared to genomic DNA from non-malignant cellular sources, with distinct fragmentation profiles observed between malignant and healthy populations[10]. DNA methylation, an early epigenetic driver of tumorigenesis, has tissue-specific and relatively high-abundance advantages[11-13]. For example, abnormal methylation of SOX17 and ZNF582 can be detected in the precancerous stage (sensitivity 41.3%, specificity 93.6%). However, traditional methylation detection methods (like MSP) have limitations in detecting low-abundance cfDNA (<0.1%) and lack whole-genome coverage, restricting clinical application.

Recent studies indicate that multi-omics integration strategies can significantly enhance detection performance through multi-dimensional feature complementarity. Combining cfDNA methylation features with protein markers boosts Gynecological Malignancies detection sensitivity to 81.9% (specificity 96.9%)[14], and integrating copy number variations and nucleosome footprints raises colorectal cancer detection AUC to 0.94[15].

Here, based on multicenter blood specimens from ESCA patients and high-risk healthy individuals, we developed an early detection model by combining cfDNA methylation patterns and fragmentation characteristics. Using LP-WGS and targeted methylation sequencing, we designed a three-layer stacked ensemble machine learning framework to integrate multi-omics genomic features. Through iterative optimization of five machine learning algorithms, this approach generated the Esophageal Cancer Screening

with Integrated Model(ESim-seq)—a stable, high-accuracy tool for ESCA screening. We assessed the separate and combined detection efficacy of fragmentomics and methylation features, providing methodological references for optimizing ESCA screening and data for future prospective cohort studies.

Results

Participants characteristics of cohort

The overview workflow of this study for the ESCA early detection model profiling is depicted in Figure 1A. For feature assessment and model development, this study enrolled 309 participants across 5 centers, including 137 ESCA patients and 172 high-risk individuals(Figure 1B). They were randomly split into a training set (N=185) and a validation set (N=124). The training cohort comprised 82 ESCA patients and 103 high-risk individuals, while the test cohort included 55 ESCA patients and 69 high-risk individuals. The proportion of ESCA patients in stages I-IV was 21.95%, 23.17%, 37.80%, and 10.98%, respectively, with most being ESCC patients. Detailed clinical characteristics are in Table 1.

cfDNA features selection and performance evaluation

Prior to developing the ESCA early-screening model, we obtained WGBS data from 31 ESCA tissues, 21 adjacent normal tissues, and 32 white blood cells. Using our self-developed CoBFP methylation block-partitioning algorithm, a methylation-specific panel covering 3,853 CpG sites (about 41.7 kb of the human genome) was designed based on the 84 samples. Evaluation showed the methylation markers could distinguish cancer from healthy samples and separate them in t-SNE analysis. We further calculated the ratios of different types of methylation signal regions, including the block methylation ratio (BMR), high methylation fragment ratio (HFR), and co-methylation fragment ratio (CFR). The distribution of these features in the cohort revealed distinct differences between cancer patients and healthy individuals.(Figure 2A-B). Unsupervised PCA of methylation features also clearly separated cancer (red) from healthy (blue) individuals (Figure 2C-E, Supplementary Figure 1A-C).

LP-WGS retains DNA genetic and fragment information to differentiate cancer from high-risk healthy individuals. We identified four cfDNA fragment-based features (CNV, FSD, FSR, NFC) that can distinguish cancer from non-cancer. The heatmap showed differences in these features between cancer and non-cancer groups (Supplementary Figure 1D-G). PCA of cfDNA features also revealed significant differences (Figure 2F-I), indicating their potential for ESCA screening.

The results have indicated methylation and cfDNA fragmentomic features can effectively distinguish ESCA from high-risk healthy individuals, providing a solid basis for model construction.

Performance of the ESCA multiomics screening model

In developing an early-screening model for ESCA, we built two prediction models. One integrates three methylation features, and the other integrates four fragment-omics features. The TMS model achieved an AUC of 0.979 (95% CI: 0.957-1.000) in the training set and 0.969 (95% CI: 0.936-1.000) in the validation set (Figure 3A-B, Supplementary Table 1). The LP-WGS model had a higher AUC of 0.995 (95% CI: 0.984-1.000) in the trainning set and the same 0.991 (95% CI: 0.973-1.000) in the validation set. . The prediction sensitivity of the TMS model and LP-WGS model reached 78.2 % (95 % CI: 69.8-86.6 %) and 81.8% (95 % CI:

74.0-89.6%) at the specificity of 99% in the validation set (Figure 3C, Supplementary Table 1).
To boost the prediction sensitivity, we developed a multimodal model named ESim-seq by combining fragment-omics and methylation features. This model demonstrated superior performance with an AUC of 0.998 (95% CI: 0.991-1.000) in the training set and 0.994 (95% CI: 0.979-1.000) in the validation set. At 99% specificity, the integrated model achieved a sensitivity of 87.3%(95% CI: 80.7--93.9%) in the validation set.
In summary, compared to single-modality models, the multimodal model (ESim-seq) which combines methylation and fragment-omics features achieved further improvement in predictive performance, showing great promise for ESCA early screening.

ESCA is linked to risk factors like hot food consumption, preserved food intake, smoking, and heavy drinking. We conducted stratified analyses across variables including gender, age, smoking status, drinking frequency, hot food consumption habits, and preserved food intake to compare the predictive abilities of single-omics models and the integrated ESim-seq model (Supplementary Figure 2, Supplementary Table 2). We assessed ESim-seq performance across tumor stages. At 99% specificity, sensitivity was stage-dependent (Figure 3D): ESCA with stage I were detected at 75% sensitivity (95% CI: 54.7-95.3%), Stage II at 70.0% (50.9-89.1%), Stage III at 96.0% (90.5-100%), and Stage IV at 100% (100-100%) in the validation set. These data demonstrate that ESim-seq maintains clinically meaningful sensitivity even in early-stage disease. Specifically, whether stratified by age, gender, drinking habits, dietary preferences, or other dimensions, the ESim-seq model offers more accurate and reliable predictions.

incidence-adjusted potential clinical benefit

In the idealized maximum interception scenario, the ESim-seq model was assessed using interception modeling to evaluate the stage distribution and survival rates. Our cohort predominantly comprised patients diagnosed at advanced stages, with low early-detection rates. Under ideal conditions, ESim-seq-based screening significantly increased the proportion of early-stage patients: 56.3% at stage I, 27.4% at stage II, 16.1% at stage III, and only 0.3% at stage IV. This indicates that ESim-seq enhances early-cancer detection, identifying more patients while their conditions are treatable.
Analysis of survival data post-ESim-seq screening revealed that 38.7% of patients died, 33.5% survived, and 27.8% were alive with tumor(Figure 4). Thus, the ESim-seq early-screening model not only improves early-cancer detection but also boosts survival rates and enhances the prognosis for ESCA patients.

Discussion

In this study, we focused on patients with ESCA and individuals at high risk of developing the disease. We innovatively developed a CoBFP-based algorithm for methylation measurement extraction. Using this algorithm, we screened for esophageal cancer-specific features in both methylation and fragmentomics, and rigorously evaluated the identified cancer-associated features. After meticulous performance comparison and screening, we selected the top-performing three methylation features and four fragmentomics features, and constructed methylation and fragmentomics-based classifiers. Furthermore, through a 3-layer Stack Ensemble model framework, we integrated the four fragmentomics features and three methylation features to develop an integrated stacked model. This sensitive of integrated model showed superior performance in distinguishing ESCA from high-risk healthy individuals compared to single-modality models, highlighting the advantages of

a multi-dimensional integration approach. Additionally, our study revealed the potential survival benefits of this integrated model under ideal conditions, offering a new perspective and a powerful tool for the early screening and intervention of ESCA.

In our preliminary work on screening methylation features, we innovatively developed a novel methylation feature screening algorithm. We introduced a new computational method called CoBFP, which is based on the consistency of methylation states among neighboring CpG sites, and performed genome-wide demarcation of co-methylation regions. Using a 3-CpG local sliding window method, we analyzed the methylation status of each CpG site at fragment level, and searched co-methylated and co-unmethylated blocks within the candidate bin. Compared to traditional differential methylation regions, CoBFP blocks can reduce methylation background noise and enhance the sensitivity of cancer early detection. The ESCA methylation data screened by this method achieved an AUC of 0.969 in the validation set, indicating excellent performance. We also applied the CoBFP algorithm to lung cancer early-screening-related studies. In these studies, the model showed an AUC of 0.973 in the training set and an AUC of 0.980 in the validation set[16].

ESCA develops slowly in most high-risk populations, progressing from squamous cell hyperplasia to dysplasia, carcinoma in situ, early-stage cancer, and invasive cancer over several years[17-18]. Early tumor detection during this period is crucial for subsequent treatment[19]. Thus, early diagnosis is vital for ESCA management. Current early-diagnosis methods for ESCA mainly include imaging, tumor-marker detection, and endoscopic screening. However, these traditional methods have limitations like invasiveness, low patient compliance, and insufficient sensitivity. In recent years, non-invasive early-screening models based on blood tests have gained attention[20], with markers like methylation[21-22], fragmentomics[23-24], microRNA(miRNA)[25-26], and mutations[27]. Among these, methylation and fragmentomics are more established and widely used in early-screening models for various cancers. They outperform miRNA and mutations in detection sensitivity and specificity, accurately indicating tumor presence. Moreover, their characteristics can be detected in the early cancer stage, giving them a unique edge in early cancer screening. In contrast, our plasma-cfDNA-based detection method is more non-invasive and sensitive, enabling earlier detection of potential tumors and buying valuable treatment time.

In the validation set of this study, the fragmentomics model achieved an sensitivity of 81.8%, compared to an sensitivity of 78.2% for the methylation model, indicating that fragmentomics outperformed methylation in terms of sensitivity. By integrating features from both fragmentomics and methylation, our multimodal model not only significantly enhanced detection performance but also corrected false-positive results that may arise from single-modality models (Supplementary Figure 3). This integrated approach has been validated in previous studies. For instance, a study combining fragmentomics with 5-hydroxymethylation features reported an AUC of 0.934 (95% CI: 0.867-1.000), with a sensitivity of 82.4% and specificity of 88.2%[28]. Moreover, in another study, the AUC for esophageal squamous cell carcinoma screening improved from 0.90 to 0.99 when using a multimodal model that combined methylation and fragmentomics[29]. Our results demonstrate that the multimodal model integrating fragmentomics and methylation features exhibits superior performance in early ESCA screening, achieving the highest AUC value among previously reported cfDNA-based early-screening studies for ESCA[29-32]. This finding provides new insights for future clinical trials on early

ESCA detection and underscores the potential of multimodal analysis in improving both detection sensitivity and specificity.

This study has several limitations. First, the limited sample size may affect the comprehensiveness and accuracy of model performance evaluation, and may cause robustness issues when distinguishing early-stage cancer from high-risk healthy individuals. Second, our study design introduces a perfect confounder between disease status and collection site, which could potentially inflate performance estimates. Third, patients with precancerous lesions of ESCA (e.g., squamous intraepithelial neoplasia) were not included in this study. Given their importance as a key target population for early ESCC detection, their absence is a significant gap. In summary, larger-scale, more diverse prospective randomized studies are urgently needed in the future to further optimize and validate the clinical efficacy of our ESCC early-detection model.

In summary, we developed an innovative method for early ESCA detection using blood-based cfDNA. By integrating advanced methylation and fragmentomics analyses, we built a multimodal model with excellent screening performance. Upon prospective validation, this multi-omics strategy, when combined with traditional endoscopy, can offer a comprehensive and precise screening approach for high-risk populations. This can significantly enhance early-detection rates and patient survival, advancing ESCA care.

Methods

Participant enrollment

This study retrospectively collected and analyzed data from two groups of participants, ESCA patients and high-risk healthy controls. The ESCA patients diagnosed by histopathological examination following the current clinical practice guidelines in China were recruited from the First Hospital of China Medical University, the National Cancer Center/Cancer Hospital of the Chinese Academy of Medical Sciences, and the First Hospital of Jilin University. All patients were newly diagnosed and had not received any form of therapy before blood collection. High-risk healthy participants were enrolled from Dandong First Hospital and Chifeng Chinese and Mongolian Medical Hospital. None of the healthy participants had a diagnosis or past history of cancer at the time of blood collection. All high-risk healthy participants were confirmed by China guideline[33]

The study protocol was approved by the Medical Science Research Ethics Committee of the First Hospital of China Medical University （2022-403-2）， the National Cancer Center/Cancer Hospital of the Chinese Academy of Medical Sciences, the First Hospital of Jilin University, Dandong First Hospital and Chifeng Chinese and Mongolian Medical Hospital. All participants provided written informed consent. All procedures were carried out in accordance with the principles of the Declaration of Helsinki.

Sample collection and cfDNA extraction

Two 10-mL tubes of peripheral blood from patients were collected using Apostle MiniMax cfDNA Blood Collection Tubes (Apostle) and were gently inverted 10 times to mix immediately. The median time between blood collection and plasma isolation was <2 days (range: 1-5 days). Plasma was separated by a two-step centrifugation process. Plasma was first separated from the peripheral blood by centrifugation at 1800 g for 10 min at 4 °C. The upper plasma was then transferred to new tubes and centrifuged at 16000 g for 15 min at 4 °C

to remove any remaining cellular debris. After centrifugation, the separated plasma was stored at -80 °C until the subsequent cfDNA extraction.

cfDNA was extracted from 8 to 10 mL of plasma using the Magbead Free-Circulating DNA Maxi Kit(CW2560S,CWBIO) on an Apostle AE2130 Nucleic Acids Extraction Automation System (CWBIO), and stored in LoBind tubes (Eppendorf AG). The concentration and quality of cfDNA were assessed using the Agilent 4200 TapeStation System (Agilent Technologies). cfDNA quantification was measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) on a Qubit 4 Fluorometer (Thermo Fisher Scientific). The extracted cfDNA was stored at -80 °C.

CfDNA methylation library construction and Targeted methylation sequencing

Methylation of cfDNA was detected using the EpiArtDNA Methylation Library Kit (EM301-02,Vazyme) via a two-step enzymatic conversion process. Briefly, 5-30 ng of extracted cfDNA was end-repaired and dA-tailed in a 60 μL reaction containing 50 μL DNA, 7 μL End Prep Buffer and 3 μL End Prep Enzyme Mix at 20 °C for 30 min, followed by 65 °C for 30 min. The cfDNA was then ligated with the adaptor at 20 °C for 15 min. After a magnetic bead purification step, 17 μL of TET2 reaction mixture and 5 μL of diluted Fe (II) solution were added to the 28 μL of purified DNA. After incubation at 37°C for 1 h, the reaction was terminated by adding 1 μL of Stop Reagent and incubated at 37°C for 30 min. The TET2 converted DNA was cleaned using 90 μL Purification Beads and eluted in 17 μL Elution Buffer. Oxidized DNA was denatured by the addition of 4 μL formamide and incubated at 85°C for 10 min. Deamination of unmethylated cytosines was performed immediately by adding 12 μL of APOBEC reaction mixture and 68 μL of nuclease-free water to the 20 μL of denatured DNA, followed by incubation at 37°C for 3 h. Deaminated DNA was then purified with 100 μL Purification Beads. Pre-library construction was performed using 20 μL of deaminated DNA, 5 μL of IDT Index Primer and 25 μL of VAHTS HiFi Amplification Mix V3(NE103-P1, Vazyme) for 7-9 cycles of PCR amplification. Pre-libraries of cfDNA were quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) on a Qubit 4 Fluorometer (Thermo Fisher Scientific). Target enrichment was performed using the Twist Fast Hybridization and Wash Kit (Twist Bioscience) according to the manufacturer's protocol. Target enrichment cfDNA methylation libraries were sequenced on the Illumina NovaSeq 6000 platform (2x150 bp) with an average sequencing depth of 1452x (449~2880).

Low-pass whole genome sequencing

10 ng of cfDNA per sample was input into library preparation using the VAHTS Universal DNA Library Prep Kit (ND610，Vazyme). Briefly, End-repair and A-tailing of cfDNA were performed using the End Repair & A-Tailing Enzyme mix, followed by an adaptor ligation to both ends. cfDNA libraries were then amplified underwent 7 PCR cycles. The final concentration of each library was determined using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and Qubit 4 Fluorometer (Thermo Fisher Scientific). Low-pass whole genome sequencing was performed using 150-bp paired-end runs (300 cycles) on the DNBSEQ-T7 platform. The average sequencing depth for each library was 3.02 x(0.46~9.6 x).

CoBFP blocks segregation

Using whole-genome bisulfite sequencing(WGBS) data from 31 esophageal carcinoma tissues, 21 adjacent normal tissues, and 32 white blood cells (WBC), we extracted the methylation status of CpG loci for each

fragment. Based on the genomic distribution of CpG loci, we split the whole genome regions into bins that if the distance of two adjacent CpG sites ≤ 500bp and filtered bins containing < 3 CpG sites. Within each bin, according to methylation status of fragments, we implemented a 3-CpG local sliding window approach to calculate the Co-Methylation/unMethylation Fragment Percent (CoBFP).

If the CoBFP score between two CpG sites within a bin was ≥0.9, the local sliding window process was continued; otherwise, the bin was split at that CpG site, and the process was restarted from the next CpG site. This process was repeated to split all bins into candidate blocks. After filtering out blocks containing <3 CpG sites, a total of 2,410,438 CoBFP blocks covering 8,813,582 CpG sites were identified for subsequent analysis(Supplementary Figure 4A).

Targeted methylation panel design

WGBS data from 84 samples were also used to identify differentially methylated CoBFP blocks. By comparing methylation levels of (1) cancerous and corresponding adjacent tissues, (2) cancerous tissues and wbc to identify cancer-specific blocks. The blocks with FDR-adjusted P value <0.001 by t-test and |△mean methylation ratio|⩾0.4 and |△median methylation ratio|⩾0.4 were considered significant, and a total of 1,499 blocks were obtained. Based on these data, a customized panel of 3,853 CpG sites, spanning ~41.7kb(41708bp) of the human genome were designed for further test(Supplementary Table 3).

Methylation data analysis and features extraction

The samples were sequenced with 150 bp paired-end reads on the Illumina platform. FASTQ files were generated from raw BCL data using bcl2fastq(v2.20.0.422). After base calling, the fastp(v0.23.0)[34] software was used to trim Illumina-specific adapters and low-quality bases for raw sequencing reads. Then We use bwameth(v0.2.2, https://github.com/brentp/bwa-meth) to align the trimmed reads to the human reference genome(GRCh37/hg19) with default parameters and remove PCR duplicates by samblaster[35], the bam files were sorted by samtools[36]. Next, The reads with Mapping Quality ⩽ 20 were removed, and the remaining paired reads were stitched to obtain the original fragment information. Extract the methylation status of CpG sites within the overlapping regions between each fragment and the block regions, then mark the methylated CpG sites as 1 and the non-methylated sites as 0 to generate a methylation binary string for each fragment. Fragments with more than 3 methylated sites with CHH and CHG contexts are considered as incompletely converted fragments and will not be analyzed further. At last, the Block Methylation Ratio (BMR), Hyper Methylation Fragment Ratio (HFR) and Co-methylation Fragment Ratio (CFR) were extracted from methylation binary string, following the protocols described in our previous study[16].

Low-pass whole genome data analysis and feature extraction

The raw sequencing data in FASTQ format were trimmed and filtered using fastp (v0.23.4)[34]. The trimmed reads were aligned to the hg19 reference genome using BWA-MEM (v0.7.17)[37]. Aligned BAM files were sorted with SAMtools (v1.17)[36], and duplicate reads were removed using Sambamba (v1.0.0)[38]. Paired-end sequencing reads were merged into single cfDNA fragments, the sequences with fragment lengths between 50-500 bp were retained for downstream analysis. To reduce GC-related biases introduced by PCR amplification during library preparation, GC correction method was applied to the sequencing data based on DELFI method[23].

Copy number variation (CNA), fragment size distribution (FSD), and fragment size ratio (FSR) features were extracted according to a previous study[16]. Nucleosomal fragment coverage (NFC) was calculated based on the Griffin method, we first generated 401 transcription factor coverage profiles, then calculated the central coverage of the nucleosome region spanning from -30 bp to +30 bp.

Stacking model development

We employed a Principal Component Correction (PCC) denoising strategy to mitigate technical batch effects and background noise before model building, as described in previous study[16, 39], We developed a 3-layer stacked ensemble model incorporating two omics datasets (LP-WGS and TMS) with seven features to achieve enhanced model performance. To prevent overfitting, a 5-fold cross-validation strategy was implemented during the model training process. The training set samples were partitioned into five equally sized datasets. In each iteration, the base models were trained on four-fifths of the data, and predictions were generated on the remaining one-fifth which was not involved in parameter estimation and model training. In this study, we utilized the Scikit-Learn library (v1.4.2) and employed five distinct supervised learning algorithms: Light Gradient Boosting Machine (LGBM), XGBoost (XGB), Random Forest Classifier (RFC), Logistic Regression Classifier (LRC) and multilayer perceptron classifier(MLPC). The ROC curve analysis (Supplementary Figure 5) demonstrated that the baseline-ensemble model consistently outperforms the majority of the single baseline models, thereby validating the efficacy of the ensemble strategy, which was consequently selected as the final modeling approach. The detailed model-building process is as follows(Supplementary Figure4B):

Layer 1: The feature matrix of two omics (LP-WGS: CNA, FSD, FSR, NFC; TMS: BMR, CFR, HFR) were separately input into five distinct machine learning classifiers. For each feature, prediction scores from five classifiers were generated.

Layer 2: The prediction score matrix from Layer 1 were used as input for the second-layer ensemble model. For each of the eight features, the LRC was applied to estimate the weight values of all base models (combinations of different features and algorithms) by minimizing the least squares error, generating a new set of predicted scores.

Layer 3: The predicted scores matrix from Layer 2 were input into the third-layer ensemble model. For each omics dataset (LP-WGS and TMS), an LRC was applied to generate integrated predicted scores at this layer.

ESim-Seq Ensemble: In the final step, the LRC was applied to the predicted scores from Layer 3. This step produced the final ESim-Seq predicted scores by combining the two omics predicted scores.

Potential clinical benefits analysis

To evaluate the potential clinical benefits of our diagnostic model, we used a phase transition model, as described by Hubbell et al[40]. We compared the outcomes of standard care versus screening, using the tumor stage distribution in our model and the 5-year survival rate for esophageal squamous cell carcinoma (ESCA) to assess the potential clinical benefits of the diagnostic model. This approach helps to identify the survival advantage that the diagnostic model may offer in clinical practice.

Statistical analysis

All statistical analyses were performed using Python (v. 3.9.18) and R (v. 4.3.0). The receiver operating characteristic (ROC) curve was analyzed using the sklearn package (v. 1.4.2). Sensitivity, specificity, accuracy,

area under the curve (AUC), and corresponding 95% confidence intervals (CI) were calculated by comparing the predictions of the cancer early-screening model with clinical diagnostic results using Python (v. 3.9.18). t-SNE analysis was carried out using R package 'Rtsne'.

Declaration statements

Data Availability

The datasets generated and/or analysed during the current study are available from corresponding author upon reasonable request.

Acknowledgments

This research was funded by Liaoning Provincial Applied Fundamentals Research Program (No. 2025JH2/101330093, W.L), Natural Science Foundation of Liaoning Provincial Science and Technology Department(2023-MS-159, W.L), Liaoning "Xingliao Yingcai Program" Medical Masters Foundation of Liaoning Provincial Health Commission(YXMJ-QN-08,W.L), the National Natural Science Foundation of China(22478441,Y.L) and Beijing-Tianjin-Hebei Basic Research Project of Tianjin Science and Technology Bureau(22JCZXJC00040,Y.L). We would like to express our sincere gratitude to all the participants and researchers involved in this study. We are particularly grateful to Professor Chuanjiang Zhou from Benxi Central Hospital, Professor Ran Xu from China Medical University Affiliated Shengjing Hospital, Professor Xueying Yang from China Medical University Affiliated Fourth Hospital, and Yutao Wei from Jining First People's Hospital for their critical insights and valuable input during the research process.

Author contributions

W.L, Y.C, J.W, Y.W and H.Y contributed equally to this work. Conceptualization, W.L and X.S; data curation, W.L and D.C; formal analysis, Y.C, J.W, Y.W and H,Y; funding acquisition, D.C, Y.L, W.L and S.X; investigation, X.X, and X.Z; methodology, Y.D, P.Z, L.Y, X.W and W.D; project administration, X.H, Z.L, L.L, W.M and M.S; software, D.Y, L.Y, X.L and Y.L; writing – original draft, C.X,Y.C, J.W, Y.W and H,Y.; writing – review and editing, D.C, W.L and S.X. All authors critically revised the paper for important intellectual content and approved the final version.

Competing interests

Authors X.X., Y.D., P.Z., L.Y., X.W., X.Z., W.D., and D.C. are employees of Jiangsu Simcere Diagnostics Co., Ltd. The remaining authors declare that they do not have any competing interests.

Reference

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 74, 229–263 (2024) .

2. Jiang, W. et al. Current status and perspectives of esophageal cancer: a comprehensive review. Cancer Commun (Lond). 45(3), 281-331 (2025) .

3. Li, SW. et al. Deep learning assists detection of esophageal cancer and precursor lesions in a prospective,

randomized controlled study. Sci Transl Med. 16(743), eadk5395 (2024) .
4. Zeng,H. et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. Lancet Glob Health. 6, e555–e567 (2018).
5. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship. Am J Surg Pathol. 40(12), e94-e102 (2016) .
6. Han, B. et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent. 4(1), 47-53 (2024) .
7. Li, H. et al. Profiles and Findings of Population-Based Esophageal Cancer Screening With Endoscopy in China: Systematic Review and Meta-analysis. JMIR Public Health Surveill. 9, e45360(2023) .
8. Alunni-Fabbroni, M. et al. Circulating DNA as prognostic biomarker in patients with advanced hepatocellular carcinoma: a translational exploratory study from the SORAMIC trial. J Transl Med. 17(1), 328(2019) .
9. Lennon, AM. et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. Science. 369(6499), eabb9601 (2020) .
10. Mouliere,F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 10:eaat4921 (2018) .
11. Hoadley, K.A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 158, 929–944 (2014) .
12. Guo, S. et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat. Genet. 49, 635–642 (2017) .
13. Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat. Commun. 11, 3475 (2020) .
14. Feng, Z, et al. Performance of Liquid Biopsy-Based Multi-Omics Biomarkers for Early Detection of Gynecological Malignancies: A Prospective Study (PERCEIVE-I). Adv Sci (Weinh). 12(20), e2401760 (2025) .
15. Gao,, YIntegration of multiomics features for blood-based early detection of colorectal cancer. Mol Cancer. 23(1), 173 (2024).
16. Xue, R. et al. Evaluation and integration of cell-free DNA signatures for detection of lung cancer. Cancer Lett. 604, 217216 (2024) .
17. Dawsey, SM. et al. Squamous esophageal histology and subsequent risk of squamous cell carcinoma of the esophagus. A prospective follow-up study from Linxian, China. Cancer. 74(6), 1686-92(1994) .
18. Vantanasiri, K, Kamboj, AK, Kisiel, JB, Iyer, PG. Advances in Screening for Barrett Esophagus and Esophageal Adenocarcinoma. Mayo Clin Proc. 99(3), 459-73 (2024).
19. Pennathur, A, Gibson, MK, Jobe, BA, Luketich, JD. Oesophageal carcinoma. Lancet. 381(9864), 400-12 (2013).
20. Rai, V, Abdo, J, Agrawal, DK. Biomarkers for Early Detection, Prognosis, and Therapeutics of Esophageal Cancers. Int J Mol Sci. 24, 3316 (2023).
21. Schrag, D. et al. Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. Lancet. 402(10409), 1251-1260 (2023).
22. Gao, Q. et al. Unintrusive multi-cancer detection by circulating cell-free DNA methylation sequencing (THUNDER): development and independent validation studies. Ann Oncol. 34(5),486-495 (2023).

23. Mathios, D,. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 12(1):5060 (2021).
24. 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(1), 129 (2022).
25. Miyoshi, J. et al. A microRNA-based liquid biopsy signature for the early detection of esophageal squamous cell carcinoma: a retrospective, prospective and multicenter study. Mol Cancer. 21(1), 44 (2022).
26. Shen, Y. et al. Identification of Novel Circulating miRNA Biomarkers for the Diagnosis of Esophageal Squamous Cell Carcinoma and Squamous Dysplasia. Cancer Epidemiol Biomarkers Prev. 28(7), 1212-20 (2019).
27. Wong, D. et al. Early Cancer Detection in Li-Fraumeni Syndrome with Cell-Free DNA. Cancer Discov. 14(1), 104-119 (2024).
28. Lu, D. et al. Plasma cell-free DNA 5-hydroxymethylcytosine and whole-genome sequencing signatures for early detection of esophageal cancer. Cell Death Dis. 14(12), 843 (2023).
29. Liu, J. et al. Multimodal analysis of cfDNA methylomes for early detecting esophageal squamous cell carcinoma and precancerous lesions. Nat Commun. 15(1), 3700(2024).
30. Jiao, Z. et al. Leveraging cfDNA fragmentomic features in a stacked ensemble model for early detection of esophageal squamous cell carcinoma. Cell Rep Med.5(8):101664 (2024).
31. Li, Y. et al. Genome-Scale Multimodal Analysis of Cell-Free DNA Whole-Methylome Sequencing for Noninvasive Esophageal Cancer Detection. JCO Precis Oncol. 8:e2400111 (2024).
32. Zhang, R. et al. A multicenter prospective clinical trial reveals cell free DNA methylation markers for early esophageal cancer. J Clin Invest. 135(8),e186816 (2025).
33. He, J. et al. China guideline for the screening, early detection and early treatment of esophageal cancer (2022, Beijing)]. Zhonghua Zhong Liu Za Zhi. Jun 23;44(6), 491-522 (2022) .
34. Chen, S, Zhou, Y, Chen, Y, Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17), i884-i890 (2018).
35. Faust, GG, Hall, IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 30(17), 2503-5 (2014) .
36. Danecek, P. et al. Twelve years of SAMtools and BCFtools. Gigascience. 10(2):giab008 (2021).
37. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv e-prints (2013).
38. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinforma. Oxf. Engl. 31, 2032–2034 (2015).
39. Lee, YH. et al. Reducing system noise in copy number data using principal components of self-self hybridizations. Proc Natl Acad Sci U S A.109(3), E103-E110 (2012) .
40. Hubbell, E., Clarke, C.A., Aravanis, A.M., and Berg, C.D. Modeled Reductions in Late-stage Cancer with a Multi-Cancer Early Detection Test. Cancer Epidemiol. Biomarkers Prev. 30, 460–468 (2021).

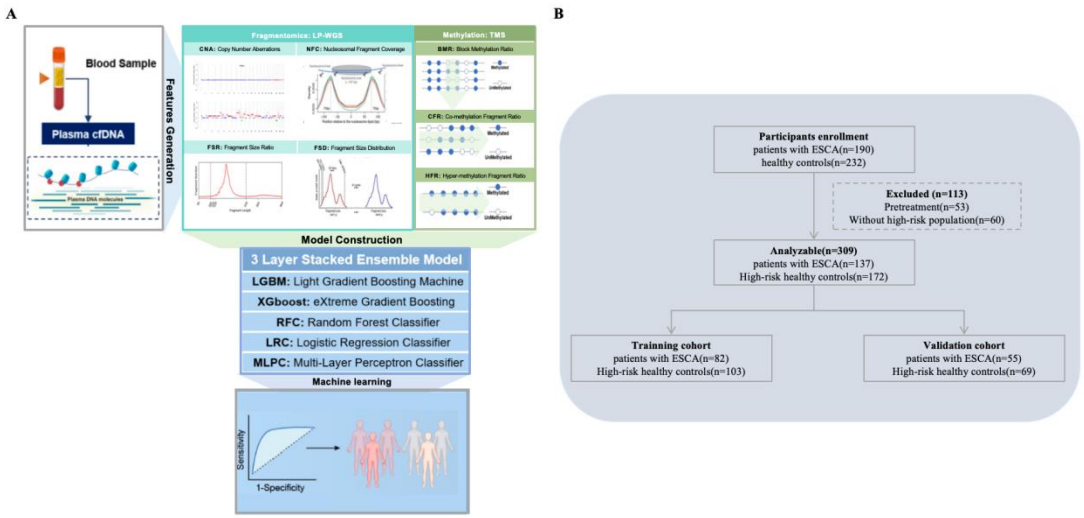

Figure 1 Study design and participant enrollment. (A) Overview of the multi-center retrospective case-control study workflow, showing data collection, multi-omics profiling (methylation and fragmentomics), and three-layer stacked ensemble model development. (B) The diagram of participant enrollment across five centers, with final numbers in training (n=137) and validation (n=172) cohort.

ESCA: Esophageal cancer; cfDNA: cell-free DNA; LP-WGS: low-pass whole-genome sequencing; TMS: targeted methylation sequencing.

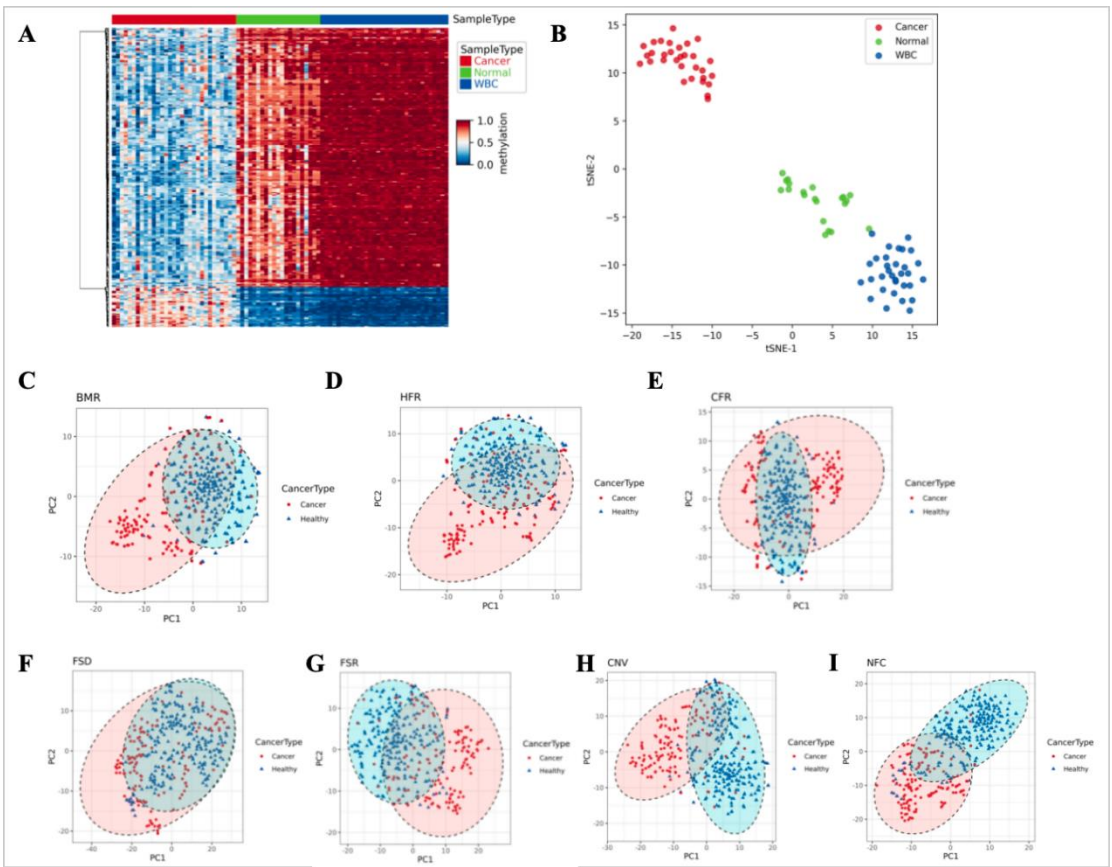

Figure 2 Features selection and performance. (A) The heatmaps display the differences of the cancer-specific methylation regions between cancers and non-cancer; (B) Visualization by t-SNE based on the cancer-specific methylation regions for cancer signal detection; (C-E) Visualization by t-SNE based on the methylation features; (F-I) Visualization by t-SNE based on the fragmentomic features.

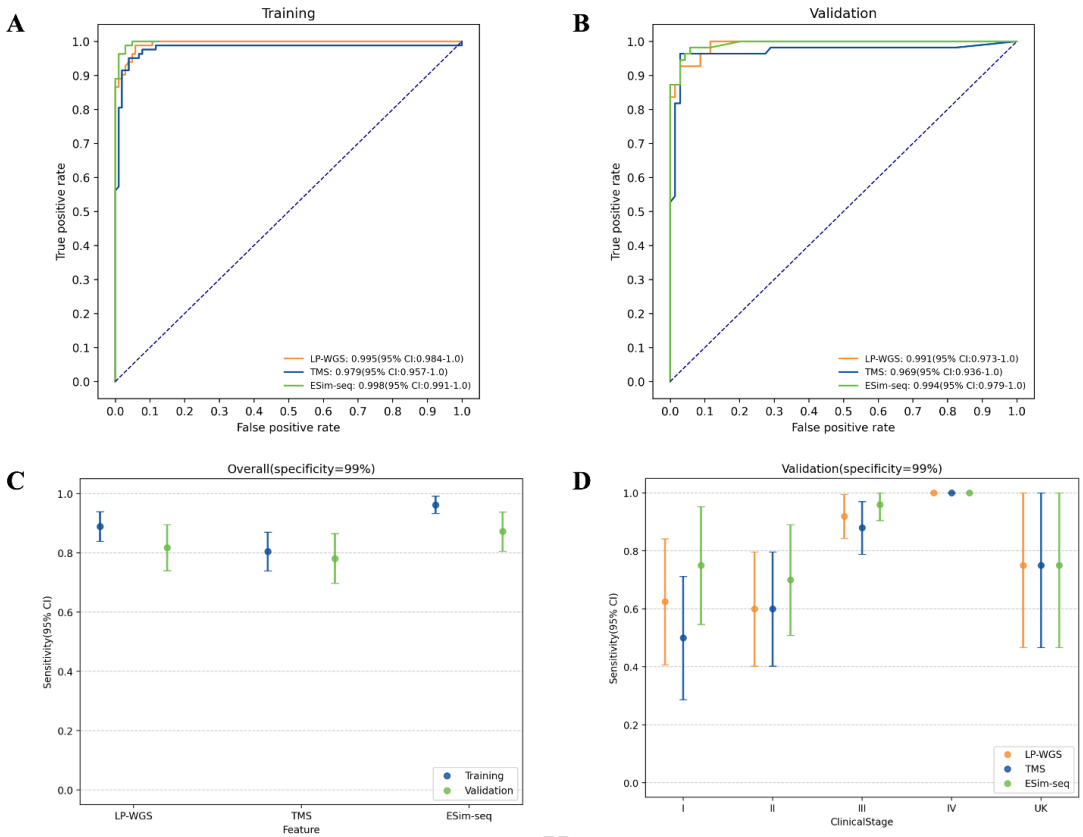

Figure 3 ROC curves and model performance metrics.（A-C）The performance of different model for ESCA patients and high-risk individuals. ROC curves and the sensitivity of the ESim-seq model in the training and validation sets. (A-B) The training and validation set ROC curves with 95% confidence intervals. (C) The sensitivity at 99% specificity for the ESim-seq model and single-omics models.Error bars represent 95% confidence intervals (Wilson score method)

AUC: area under the curve; CI: confidence interval.

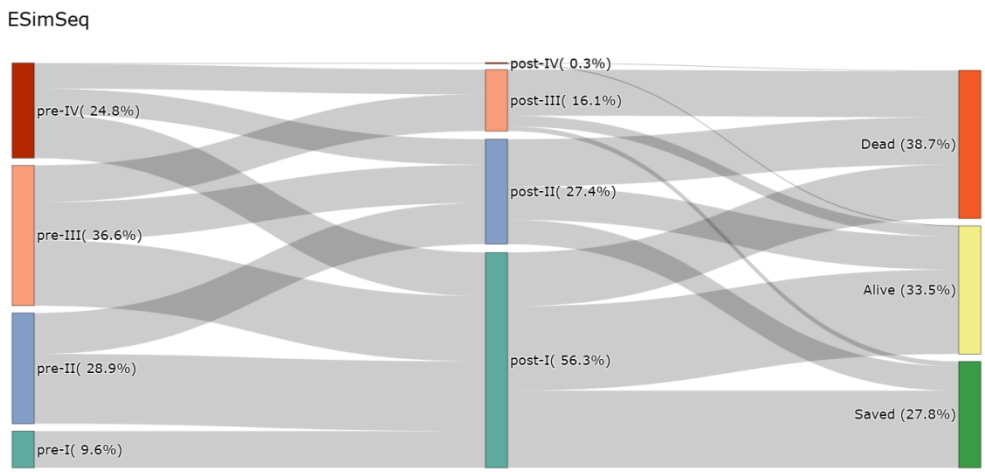

Figure 4 Potential clinical benefit with stage and mortality of the ESim-seq models

Table 1 Participant characteristics of the study

Variable	Trainning set		Validation set	
	Cancer	Non-Cancer	Cancer	Non-Cancer
Total, (n)	82	103	55	69
Age, Mean±SD	63±8	56±8	63±8	55±8
Sex, (n, %)				
Male	77(93.90)	47(45.63)	48(87.27)	30(43.44)
Female	3(3.66)	56(54.37)	6(10.91)	39(56.52)
UK	2(2.44)	0(0)	1(1.82)	0(0)
Stage, (n, %)				
I	18(21.95%)		8(14.55%)	
II	19(23.17%)		10(18.18%)	
III	31(37.80%)	-	25(45.45%)	-
IV	9(10.98%)		8(14.55%)	
UK	5(5.10%)		4(7.27%)	
Sample type(n)				
ESCC	74	-	48	-
Others	8		7	
High-risk individuals.	-	103	-	69