Beta-Validation of a Non-Invasive Method for Simultaneous Detection of Early-Stage Female-Specific Cancers

Sudhakar Ramamoorthy[1], Saranya Sundaramoorthy[2], Ankur Gupta[3,4], Zaved Siddiqui[3,4], Ganga Sagar[3], Kanury V.S. Rao[3,4], Najmuddin Saquib[3,4*] & Surapeni Krishna Mohan[5*]

[1]Department of Pathology

[2]Department of Biochemistry

[5]Department of Biochemistry, Clinical Skills & Simulations Research

Panimalar Medical College Hospital & Research Institute

Varadharajpuram, Poonamallee

Chennai – 600 123

India

[3]PredOmix Technologies Private Limited

Tower B, SAS Tower

Medicity, Sector – 38

Gurugram – 122002

India

[4]PredOmix Health Sciences Private Limited
10 Anson Road
#22-02 International Plaza
Singapore 079903

[*]Address correspondence and requests to SKM (email: krishnamohan.surapaneni@gmail.com), or NS (email: saquib@predomix.com)

Tel: 0124-4307268

ABSTRACT

Objectives: In this report, we assessed the accuracy of our previously developed method for simultaneous diagnosis of the four female cancers of the Breast, Endometrium, Cervix, and Ovary, in a clinical set-up with blinded protocol.

Materials and Methods: Our test protocol combined global serum metabolome profiling wherein data was analyzed with machine learning algorithms to extract metabolite signatures that correlated with early-stage cancers. High-resolution mass spectrometry was employed to profile the serum metabolome and the resulting data were subjected to a pre-processing pipeline to obtain the data set. The data was then analyzed using artificial intelligence algorithms to identify early-stage cancer metabolic signatures.

Results: Overall, a total of 1000 blinded samples were analyzed by generating the serum metabolome profiles, followed by sequential algorithms for cancer detection and multiclass cancer type identification. Of these 1000 samples, 797 were identified as cancer positive, while, 203 samples were identified as cancer-negative. The multiclass algorithm was then applied to the 797 cancer-positive samples, to distinguish between samples that were from patients with either endometrial, breast, cervical, or ovarian cancer. After completion of the analysis, the sample code was broken to estimate the accuracy of the results. Concerning the identification of samples that were cancer-positive, the sensitivity obtained was 99.6% whereas the specificity was 100%. For the second stage of analysis which involves 'tissue of origin', all 107 breast cancer samples were correctly identified without any false calls. The accuracy for identification of cervical and ovarian cancers was between 95-96% for each, whereas 91% for endometrial cancer.

Conclusions: Our present study validates the performance of our method for the early-stage detection of female-specific cancers in a clinical setting. Importantly, the algorithms for cancer detection and 'tissue of origin' prediction, which were initially trained using samples from Caucasian patients, retained the accuracy on samples from Indian women patients. This suggests that the performance of these algorithms was minimally influenced by variables such as ethnicity and race. Present results, therefore, also underscore the potential clinical utility of our method for early-stage diagnosis of cancers that are specific to females.

Keywords: Breast Cancer, Endometrial Cancer, Cervical Cancer, Ovarian Cancer, Untargeted Metabolomics, Machine Learning, Tissue of Origin (TOO)

INTRODUCTION

Worldwide over 6 million new cancer cases and close to 4 million cancer related deaths are reported every year (1,2). The most prominent female-specific cancers are those of the breast, cervix, endometrium, and ovary (1,3). Breast cancer is the most common and represent ¼ th of cancers in women. (4), whereas cancer of the endometrium and ovary account for 5% and 4% respectively (5,6). Cervical cancer ranks fourth with an estimated 500,000 cases globally (7).

These cancers are often detected at their more advanced stages of progression. Consequently, prognosis is usually poor (2,8,9). While treatment outcomes are significantly improved when the cancer is detected at an early stage (10), effective screening paradigms for early-stage female-specific cancers are either not currently available or suffer from inherent limitations that restrict their scope (11, 12). For example, while mammography is the recommended method for breast cancer screening, it is not a perfect test because its overall sensitivity ranges only from 75% to 85% (13). The sensitivity further decreases sharply in the case of dense breast parenchyma (14). The net consequence of these limitations is that many women with breast cancer are missed (14). Another downside to mammography is its relatively high false-positivity rate, which is further enhanced in women below the age of 40 years. As a result, the likelihood that a woman will receive at least 1 false-positive call after 10 yearly mammograms is found to be as high as 61% (15). It follows, therefore, that an improved breast cancer screening method with high sensitivity and specificity across all age groups of women is clearly needed to minimize the harm to benefit ratio and, thereby, obtain better outcomes. Efforts to achieve this are presently ongoing (16)

Current guidelines recommend three primary screening options for cervical cancer. These are cytological testing alone, standalone high-risk human papilloma virus (hrHPV) testing, and co-testing with the combination of cytological and hrHPV testing (17). While cytological testing is more widely accepted as the primary screening method for cervical cancer (18), it suffers from poor sensitivity (47-70%) although the specificity is high (19). In contrast, hrHPV standalone testing yields higher sensitivity (80%), but with a higher false positivity rate of about 15% (20,21). While sensitivity can be improved upon co-testing with both methods this, however, was shown to occur at the cost of lower specificities (20). In contrast to at least some screening methods being available for breast and cervical cancer, no such routine or standard screening test is currently available for either endometrial or ovarian cancer. Indeed, most ovarian cancer patients do not experience symptoms until its later stages. The only recourse

currently available is to either measure levels of the relatively non-specific tumor marker CA 125 in the blood, or to perform a pelvic ultrasound. However, because of poor sensitivity especially for early-stage cancer, and a high false-positivity rate, neither of these tests currently qualify as standalone screening tests for ovarian cancer (22,23). These collective results clearly highlight the need for developing more accurate and reliable methods for screening of each of these cancers. The ideal solution here would be a non-invasive test that would enable more effective surveillance of at least the more vulnerable segments of the female population.

In an earlier report we had described the development of a method for the concomitant early-stage detection of these female-specific cancers with very high accuracy. Our protocol combined untargeted metabolomics with machine learning based data analysis to extract metabolite patterns that define the early-stage cancers (24). The accuracy of detection obtained across all four cancers was 98% while a subsequent analysis also helped to identify the cancer type of the individual cancer-positive samples with reasonable accuracy. For the development of this test, we had employed stored serum samples that were obtained from biobanks. To evaluate its potential for clinical use, however, it was important for us to ascertain its performance with samples that were directly obtained from women patients in a clinical setting. The present report describes our efforts in this direction. Results obtained establish that the fidelity of both cancer detection and prediction of the tissue of origin (TOO) remained exceptionally high upon analysis of serum samples collected directly from women volunteers. Importantly, we were also able to obtain beta-validation of our method by confirming its accuracy in a blinded study protocol.

MATERIALS AND METHODS

Study design and sample collection

Blood, from both cancer patients and normal female volunteers, were collected at the Panimalar Medical College Hospital and Research Institute after first obtaining approval from the Panimalar Medical College Hospital & Research institute - Institutional Human Ethics Committee (Protocol No. PMCHRI-IHEC-034). Serum was subsequently prepared, and the samples were anonymized before submitting to the analytical team (AG, GS, ZS, KVSR, NS) for analysis.

Sample Accessioning.

Unique identity number (identifier) issued to each sample post arrival to lab. These identifiers were used for queries related to sample handling, tasks, and results. Samples were stored frozen (-80°C) until required.

Metabolite extraction.

Extraction of metabolite was done as earlier reported by Gupta *et.al* (24). Briefly, serum samples were thawed at 4C on ice. 10µl of each sample was taken and metabolites were extracted and processed for UHPLC-MS/MS as previously described (24).

Liquid Chromatography coupled with high resolution Mass Spectroscopy.

The detailed method of untargeted metabolomics was previously reported by Gupta *et.al* (24). Briefly, a Dionex LC system was employed for the high-resolution separation of serum metabolite in normal and cancer patients.

10ul of each sample was resolved by UHPLC prior to MS/MS analysis as previously described (24).

Data processing

Data generated by mass spectrometers was pre-processed through the steps of mass error correction, ion filtering, and normalisation as described earlier by Gupta *et.al* (24).

AI modeling of the data:

Detailed explanation of the method was shown by Ankur *et.al* (24). Briefly, a layered approach was used, where we first distinguish cancer from normal and then between each of the cancers. The scheme employed for training, testing and validation along with method used for determining sensitivity and specificity was as previously described (24).

RESULTS

Study rationale and the samples employed

In our previous study (24) we had developed a method, by integrating untargeted serum metabolomics by mass spectrometry and analysis of the resulting data with a machine learning algorithm to diagnose Stage 0/I of female cancers with high fidelity. These were breast, endometrial, cervical and ovarian cancers. Our protocol involved a sequential approach in which the cancer-positive samples were first distinguished from the non-cancer samples. Subsequent to this, multi-class algorithm was employed to differentiate between each of the separate cancers of the cancer-positive subset. The high accuracy obtained in this study suggested the possibility that this approach could indeed be further developed as a screening tool for the early-stage female-specific cancers.

Although our results were extremely promising, we recognized that our analysis algorithms had been developed using serum samples obtained from commercial biobanks (24). Therefore, to further explore its utility, it was necessary to evaluate the performance of our test under more relevant conditions where samples were directly obtained from patients in a clinical setting. Another important point here was that, for the algorithm development, we had employed metabolome profiles of sera that had been primarily derived from Caucasian patients. Therefore, since the metabolome composition is known to vary across different populations (28, 29), it was also relevant to establish whether these algorithms were equally effective when samples obtained from Indian women patients were tested. This question was especially pertinent given that our objective was to identify metabolite signatures that specifically characterized the disease state of interest, independent of other variables such as age, ethnicity, race, etc. We, therefore, undertook the present exercise to ascertain the accuracy of our method (24) when tested on samples directly obtained from Indian women patients.

Verification of algorithm performance for identification of Indian women patients.

Our first objective was to test the performance accuracy of our earlier developed algorithms, which were developed using data largely generated from Caucasian samples, in Indian patients. For this we used a set of 300 samples collected at the Panimalar Medical College Hospital and Research Institute (PMCHRI). Of these, 71 samples each were from patients with endometrial, cervical, or ovarian cancer, where the remaining 87 samples were from breast cancer patients. In addition, we also included 200 samples from normal volunteers. The distribution across age groups, BMI status, and cancer stages are given in Table-1.

Metabolites were extracted from each of the samples and then analyzed as previously described (24). Range of the unique metabolite numbers detected in normal control samples and the individual cancers, across the individual age groups, is depicted in Figure 1. The data was subsequently processed using our previously described in-house pipeline to eventually extract the matrix consisting of 2764 features (24).

The PLSDA plot generated using the matrix is shown in Figure 2. It is evident from this figure that the cancer samples could be unambiguously separated from the normal controls. To differentiate the cancer samples from normal controls, we applied our previously developed AI model (24) on the present data as the test set. As described in our earlier report (24), the model applies a logistic regression function to separate the female-specific cancer samples from the control group. After training the algorithm calculates a score for the individual samples by using the following formula:

$$y\_score=x0+x1*I_1+ x2*I_2+ x3*I_3+\ldots\ldots +x2764*I_{2764}$$

Here, x0 is a constant number, I_i ($1<=i<=2764$) is the intensity of metabolite i present in the respective sample.

The schematic of the approach employed is illustrated in Figure 3 and Figure 4A shows the results obtained after applying the model on the test set (24). The scatter diagram gives the scores generated by the model for control and the cancer cases. It is evident from the figure that the scores for the normal control samples are clearly distinguished from those for the cancer set. A cut-off score of 5 successfully differentiated the two sets as depicted in Figure 4B. Sensitivity, Specificity, and Accuracy all corresponded to 100%.

Differentiating between Breast, Endometrial, Cervical and Ovarian cancer samples.

Next, to distinguish the specific cancer types among the positive samples identified in Figure 4, we layered the multiclass AI model over the model described in this figure. The multiclass AI model was also developed in our earlier study (24). It is a one versus rest (OVR) classifier model, which gives four scores for each sample. Each of these scores define the likelihood of a given sample belonging to anyone of the four classes. As previously described (24), the formulae used by the algorithm to calculate the four scores for each of the sample is as follows:

$$y\_score1=y0+y1*I_1+ y2*I_2+ y3*I_3+\ldots\ldots +y2764*I_{2764}$$

$$y\_score2=z0+z1*I_1+ z2*I_2+ z3*I_3+\ldots\ldots +z2764*I_{2764}$$

$$y\_score3=a0+a1*I_1+ a2*I_2+ a3*I_3+\ldots\ldots +a2764*I_{2764}$$

$$y\_score4=b0+b1*I_1+ b2*I_2+ b3*I_3+\ldots\ldots +b2764*I_{2764}$$

y0, z0, a0, b0 represent constant numbers, I_i (1<=i<=2764) is the relative concentration of metabolite i in the respective sample.

To assess the performance of our multiclass model in terms of differentiating between the individual cancer types for Indian patients, as also from normal controls, the scores obtained from the model were plotted. In Figure 5A, we plotted the score obtained for Endometrial cancer samples against that of the group comprising of the Breast, Cervical and Ovarian (BCO) cancers. A clear distinction between Endometrial and BCO Cancer samples is seen (Fig. 5A). Control samples were also included in this analysis and sensitivity, specificity and accuracy obtained were 97%, 94%, and 95% respectively.

Next, we plotted the Breast cancer sample scores versus those of the group comprised of Endometrial, Cervical and Ovarian (ECO) cancer samples (Fig. 5B). Here again Breast Cancer could be clearly distinguished from ECO cancer sample set as shown (Fig. 5B). The sensitivity, specificity and accuracy obtained were 94%, 93.8%, and 93.9% respectively.

Figure 5C depicts the confusion matrix obtained for an analysis of Cervical cancer sample scores versus that for the group of Endometrial, Breast and Ovarian (EBO) cancer samples. Sensitivity, specificity and accuracy obtained were 92%, 97%, and 96% respectively. And, finally, Figure 5D shows the confusion matrix obtained for discriminating Ovarian cancer samples from the group comprising of Endometrial, Breast and Cervical (EBC) Cancer samples. Here, the calculated the sensitivity, specificity and accuracy were 98.5%, 97.7%, and 97.6% respectively.

Beta-validation of our algorithms through a blinded study protocol.

The results obtained so far establish that the algorithms that we had previously developed, using data primarily generated with samples obtained from Caucasians (24), were equally effective for screening of female-specific cancers in Indian women. Indeed the high sensitivity and specificity obtained, for both detection of cancer positivity as well as for identification of cancer type, also suggest that this approach is likely to be at least less sensitive to population-related differences in the subjects. To further verify the potential clinical utility of our test platform, we sought to beta-validate it by employing a blinded protocol. In this, coded samples we sent by the clinical team (SR, SS, SKM) to the team involved in sample analysis and diagnosis of the cancer state and type (analytics team, AG, GS, ZS, NS, KVSR). Here, all

public health information (PHI) identifiers/variable were protected from the latter team. Apart from sample IDs, no information on either cancer status or cancer type were provided.

A total of 1000 samples were provided by the clinical team, which the analytical team then analysed by generating the corresponding serum metabolome profiles, followed by sequential application of the cancer detection and multiclass cancer type identification algorithms. Of these 1000 samples, 797 were identified as cancer positive, whereas the remaining 203 samples were identified as cancer negative. Following this, the multiclass algorithm was then applied to the 797 samples putatively identified as cancer positive, in order to distinguish those samples that were patients with either endometrial, breast, cervical, or ovarian cancer. The cumulative results were then sent to the clinical team, which then broke the code and summarized their findings. These findings are presented in Tables 2 and 3 where Table 2 provides the results generated with the first algorithm for identification of cancer positive samples. It is evident from this table that a very high detection accuracy was obtained, with no false positives and only 3/800 true positive samples being missed as false negatives. The sensitivity obtained here was 99.6%, with a specificity of 100% (Table 2). Table 3 presents the results obtained after application of the multiclass algorithm for distinguishing between samples from patients with either breast, endometrial, cervical, or ovarian cancer. Here, the algorithm was only applied to the 797 blinded samples that were identified as cancer positive in Table 2. The results in Table 3 clearly underscore the high degree of fidelity with which the individual cancer types could be distinguished. While all 107 of the breast cancer samples were correctly identified, the accuracy for identification of cervical and ovarian cancer samples was between 95 – 96%. Endometrial cancer identification was, however, somewhat lower with an accuracy of close to 91% (Table 3). These results, therefore, provide strong validation for our method, which integrates untargeted metabolome profiling with AI-driven data analytics, for concurrent detection of female-specific cancers.

DISCUSSION

While there is an increased emphasis on detection of cancers at an early stage, recent years are also witnessing a paradigm shift towards multiple cancer detection with a single blood draw (30). These latter efforts have largely focused on analysis of circulating tumour cell-free DNA (ctDNA) by using genomic technologies and machine learning to not only detect multiple cancers simultaneously but also identify the tissue of signal origin (TOO) (31-35). The promise that such multi-cancer detection (MCD) tests hold is that they would permit detection of several cancers that otherwise go undetected until they reach an advanced stage

(36). Indeed, mathematical modelling of the potential public health impact of including multi-cancer detection tests to usual care suggested a marked positive effect with a substantial reduction in overall cancer mortality (37).

Despite the promise shown by MCDs based on ctDNA analysis, concerns have been raised that ctDNA may be a less than satisfactory proxy for liquid biopsies of tumour tissues for early detection because of limitations in both sensitivity and specificity (38, 39). In this context, the low concentration of ctDNA and its variability based on type and status of the tumour also contributes as a complicating factor (40). Furthermore, the low to negligible concentration of ctDNA present in the early stages of cancer also limits the sensitivity that can be achieved for the stages (41, 42). This concern is borne out by more recent results from clinical trials where sensitivity of detection of early-stage cancers was indeed found to be low (34, 35). Given these potential limitations with ctDNA, we preferred to test the utility of metabolomics for early-stage cancer detection. The rationale here was based on the now widely accepted notion that metabolites serve as proximal reporters of disease since their relative abundances are often directly related to pathogenic mechanisms (43-45). We hypothesized that metabolomics should be an especially relevant technique for cancer detection since cancers have significantly altered metabolism (46, 47). Therefore, the spectrum of metabolites produced can possibly yield signatures characteristic of the presence of cancer.

As demonstrated in our previous report (24) and the results described here, our expectation was indeed borne out. At least in the context of the four female-specific cancers, serum metabolome analysis coupled with AI analysis yielded a very accurate method for their simultaneous detection at the early stage, and subsequent of the tissue of origin (TOO). Indeed the high sensitivity obtained for detection of Stage-I of all the four cancers is particularly notable and unmatched by that achieved through analysis of ctDNA (34, 35). We do, however, acknowledge that our current method only enables simultaneous detection of the four female-specific cancers. Nonetheless, an important highlight of our present study is that it validates the performance of our method in a clinical setting. Particularly notable here is the fact that the algorithms for cancer detection and TOO prediction, which were trained primarily on samples from Caucasian patients (24), retained their accuracy with samples from Indian women patients. This indicated that, at least with respect to the two population groups studied, variables such as ethnicity and race did not affect performance of either algorithm. Indeed,

the high sensitivity and specificity of >99% obtained for detection of the cancer positive samples, and the high TOO prediction accuracy, in the blinded study especially underscore this, while also serving to validate the method developed by us. We do recognize, however, that a more extensive clinical trial would be required to demonstrate its clinical utility. Furthermore, it will also be of interest to explore whether early detection of additional cancers can be brought within the ambit of this approach.

REFERENCES

1. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. Cancer Epidemiol Biomarkers Prev. 2017 Apr;26(4):444-457. doi: 10.1158/1055-9965.
2. Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. Cancer Statistics, 2022. *CA Cancer J Clin.*, 2022 Jan;72(1):7-33. doi: 10.3322/caac.21708
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660.
4. Ghoncheh M, Pournamdar Z, Salehiniya H. Incidence and Mortality and Epidemiology of Breast Cancer in the World. Asian Pac J Cancer Prev. 2016;17(S3):43-6. doi: 10.7314/apjcp.2016.17.s3.43.
5. Zhang S, Gong T, Liu F, Jiang Y, Sun H, Ma X, et al. (2019). Global, Regional, and National Burden of Endometrial Cancer, 1990–2017: Results From the Global Burden of Disease Study, 2017. Frontiers in Oncology, 9. https://doi.org/10.3389/fonc.2019.01440
6. Momenimovahed Z, Tiznobaik A, Taheri S, Salehiniya H. Ovarian cancer in the world: epidemiology and risk factors. Int J Womens Health. 2019 Apr 30;11:287-299. doi: 10.2147/IJWH.S197604.
7. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health. 2020 Feb;8(2):e191-e203. doi: 10.1016/S2214-109X(19)30482-6.
8. Howlader, N. et al. (eds) SEER *Cancer Statistics Review*, 1975–2014 (National Cancer Institute, 2017).
9. Ahlquist DA. Universal cancer screening: revolutionary, rational, and realizable. NPJ Precis Oncol. 2018 Oct 29;2:23. doi: 10.1038/s41698-018-0066-x.

10. World Health Organization. Guide to early cancer diagnosis. https:// apps. who. int/ iris/ bitst ream/ handle/ 10665/ 254500/ 9789241511 940- eng. pdf? seque nce= 1& isAll owed=y (2017).
11. Pinsky PF, Prorok PC, Kramer BS. Prostate Cancer Screening - A Perspective on the Current State of the Evidence. N Engl J Med. 2017 Mar 30;376(13):1285-1289. doi: 10.1056/NEJMsb1616281.
12. Subramanian S, Klosterman M, Amonkar MM, Hunt TL. Adherence with colorectal cancer screening guidelines: a review. Prev Med. 2004 May;38(5):536-50. doi: 10.1016/j.ypmed.2003.12.011.
13. Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. Cochrane Database Syst Rev. 2013 Jun 4;2013(6):CD001877. doi: 10.1002/14651858.CD001877.pub5.
14. Chen HL, Zhou JQ, Chen Q, Deng YC. Comparison of the sensitivity of mammography, ultrasound, magnetic resonance imaging and combinations of these imaging modalities for the detection of small (≤2cm) breast cancer. Medicine (Baltimore). 2021 Jul 2;100(26):e26531. doi: 10.1097/MD.0000000000026531.
15. Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. Ann Intern Med. 2011 Oct 18;155(8):481-92. doi: 10.7326/0003-4819-155-8-201110180-00004. Erratum in: Ann Intern Med. 2014 May 6;160(9):658.
16. Gastounioti, Aimilia & Desai, Shyam & Ahluwalia, Vinayak & Conant, Emily & Kontos, Despina. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. Breast Cancer Research. (2022). 24. 10.1186/s13058-022-01509-z.
17. Terasawa T, Hosono S, Sasaki S, Hoshi K, Hamashima Y, Katayama T, et al. Comparative accuracy of cervical cancer screening strategies in healthy asymptomatic women: a systematic review and network meta-analysis. Sci Rep. 2022 Jan 7;12(1):94. doi: 10.1038/s41598-021-04201-y.
18. Nkwabong E, Laure Bessi Badjan I, Sando Z. Pap smear accuracy for the diagnosis of cervical precancerous lesions. Trop Doct. 2019 Jan;49(1):34-39. doi: 10.1177/0049475518798532. Epub 2018 Sep 15.

19. Boone JD, Erickson BK, Huh WK. New insights into cervical cancer screening. J Gynecol Oncol. 2012 Oct;23(4):282-7. doi: 10.3802/jgo.2012.23.4.282. Epub 2012 Sep 19.
20. Liang, Linda & Einzmann, Thomas & Franzen, Arno & Schwarzer, Katja & Schauberger, Gunther & Schriefer, Dirk & Radde, et al. Cervical Cancer Screening: Comparison of Conventional Pap Smear Test, Liquid-Based Cytology, and Human Papillomavirus Testing as Stand-alone or Cotesting Strategies. Cancer Epidemiology Biomarkers & Prevention. (2020). 30. cebp.1003.2020. 10.1158/1055-9965.EPI-20-1003.
21. Longatto-Filho A, Naud P, Derchain SF, Roteli-Martins C, Tatti S, Hammes LS, et al. Performance characteristics of Pap test, VIA, VILI, HR-HPV testing, cervicography, and colposcopy in diagnosis of significant cervical pathology. Virchows Arch. 2012 Jun;460(6):577-85. doi: 10.1007/s00428-012-1242-y. Epub 2012 May 5.
22. Menon U, Gentry-Maharaj A, Burnell M, Singh N, Ryan A, Karpinskyj C, et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2021 Jun 5;397(10290):2182-2193. doi: 10.1016/S0140-6736(21)00731-5. Epub 2021 May 12.
23. Henderson JT, Webber EM, Sawaya GF. Screening for Ovarian Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA. 2018 Feb 13;319(6):595-606. doi: 10.1001/jama.2017.21421.
24. Gupta, A., Sagar, G., Siddiqui, Z., Rao, K. V., Nayak, S., Saquib, N., & Anand, R. (2022). A non-invasive method for concurrent detection of early-stage women-specific cancers. Scientific Reports, 12(1), 1-12. https://doi.org/10.1038/s41598-022-06274-9
25. Brochu F, Plante P, Drouin A, Gagnon D, Richard, D, Durocher F, et al. Mass spectra alignment using virtual lock-masses. Scientific Reports, (2019). 9(1), 1-15. https://doi.org/10.1038/s41598-019-44923-8
26. Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, Gronwald W. State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics. 2012 Jun;8(Suppl 1):146-160. doi: 10.1007/s11306-011-0350-z. Epub 2011 Aug 12.
27. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites. 2014 Jun 16;4(2):433-52. doi: 10.3390/metabo4020433.

28. Lau CH.E, Siskos A.P, Maitre L. Determinants of the urinary and serum metabolome in children from six European populations. BMC Med 16, 202 (2018). https://doi.org/10.1186/s12916-018-1190-8
29. Tkachev A, Stepanova V, Zhang L, Khrameeva E, Zubkov D, Giavalisco P, Khaitovich P. Differences in lipidome and metabolome organization of prefrontal cortex among human populations. Sci Rep. 2019 Dec 4;9(1):18348. doi: 10.1038/s41598-019-53762-6.
30. Liu MC. Transforming the landscape of early cancer detection using blood tests-Commentary on current methodologies and future prospects. Br J Cancer. 2021 Apr;124(9):1475-1477. doi: 10.1038/s41416-020-01223-7. Epub 2021 Feb 9.
31. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018 Feb 23;359(6378):926-930. doi: 10.1126/science.aar3247. Epub 2018 Jan 18.
32. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV; CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020 Jun;31(6):745-759. doi: 10.1016/j.annonc.2020.02.011. Epub 2020 Mar 30.
33. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 2020 Jul 21;11(1):3475. doi: 10.1038/s41467-020-17316-z.
34. Klein, E. A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., *et. al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals Oncol.* (2021) **32**, P1167-1177
35. Lennon AM, Buchanan AH, Kinde I, Warren A, Honushefsky A, Cohain AT, Ledbetter DH, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. Science. 2020 Jul 3;369(6499):eabb9601. doi: 10.1126/science.abb9601. Epub 2020 Apr 28.
36. Ahlquist DA. Universal cancer screening: revolutionary, rational, and realizable. NPJ Precis Oncol. 2018 Oct 29;2:23. doi: 10.1038/s41698-018-0066-x.
37. Earl Hubbell, Christina A. Clarke, Alexander M. Aravanis, Christine D. Berg; Modeled Reductions in Late-stage Cancer with a Multi-Cancer Early Detection Test. Cancer Epidemiol Biomarkers Prev 1 March 2021; 30 (3): 460–468. https://doi.org/10.1158/1055-9965.EPI-20-1134

38. Fiala C, Diamandis EP. Circulating tumor DNA (ctDNA) is not a good proxy for liquid biopsies of tumor tissues for early detection. Clin Chem Lab Med. 2020 Sep 25;58(10):1651-1653. doi: 10.1515/cclm-2020-0083.
39. Keller L, Belloum Y, Wikman H, Pantel K. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. Br J Cancer. 2021 Jan;124(2):345-358. doi: 10.1038/s41416-020-01047-5. Epub 2020 Sep 24.
40. Yasai H. Challenges in circulating tumor DNA analysis for cancer diagnosis. J Nanomed Res. 2018;972):76 – 80. DOI: 10.15406/jnmr.2018.07.00180
41. Fiala, C., Diamandis, E.P. Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. BMC Med 16, 166 (2018). https://doi.org/10.1186/s12916-018-1157-9
42. Molparia, B., Nichani, E., & Torkamani, A. Assessment of circulating copy number variant detection for cancer screening. PLOS ONE, (2017) 12(7), e0180647. https://doi.org/10.1371/journal.pone.0180647
43. Shao, Y., Le, W. Recent advances and perspectives of metabolomics-based investigations in Parkinson's disease. Mol Neurodegeneration 14, 3 (2019). https://doi.org/10.1186/s13024-018-0304-2
44. Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M, et al. for "Precision Medicine and Pharmacometabolomics Task Group"-Metabolomics Society Initiative. Metabolomics enables precision medicine: "A White Paper, Community Perspective". Metabolomics. 2016;12(10):149. doi: 10.1007/s11306-016-1094-6. Epub 2016 Sep 2.
45. Puchades-Carrasco L, Pineda-Lucena A. Metabolomics Applications in Precision Medicine: An Oncological Perspective. Curr Top Med Chem. 2017;17(24):2740-2751. doi: 10.2174/1568026617666170707120034.
46. Beger RD. A review of applications of metabolomics in cancer. Metabolites. 2013 Jul 5;3(3):552-74. doi: 10.3390/metabo3030552.
47. Wang L, Liu X, Yang Q (2018) Application of Metabolomics in Cancer Research: As a Powerful Tool to Screen Biomarker for Diagnosis, Monitoring and Prognosis of Cancer. Biomark J. 4:12. doi: 10.21767/2472-1646.100050.

AUTHOR CONTRIBUTIONS

S.R, S.S., and S.K.M. coordinated the procurement of patient samples and confirmation of their diagnosis. N.S., G.S., and Z.S. led all experimental aspects of the study while data analysis was performed by A.G. The study was jointly supervised by K.V.S.R. and S.K.M. N.S., K.V.S.R., and S.K.M. wrote the paper.

COMPETING FINANCIAL INTERESTS.

A.G, G.S., Z.S. and N.S. are fulltime employees of PredOmix Technologies Private Limited. K.V.S.R. is a cofounder and owns stock in both PredOmix Technologies Private Limited and PredOmix Health Sciences Pte. Ltd. A.G., Z.S., and NS. own stock in PredOmix Health Sciences Pte. Ltd.

Table 1: Demographic, BMI, and Cancer-stage grouping of the samples employed for algorithm verification.

Sample Clinical Info						
		Clinical Status				
Parameter	**Intervals**	**Normal control** (n=200)	**Endometrial cancer** (n=71)	**Breast cancer** (n=87)	**Cervical cancer** (n=71)	**Ovarian cancer** (n=71)
Age(years)	20-30	102	0	8	1	5
	31-40	51	2	26	5	17
	41-50	29	17	19	7	27
	51-60	14	35	28	24	16
	61-70	4	13	3	20	5
	71-80	0	4	3	13	0
	81-90	0	0	0	1	1
BMI (kg/m2)	10 to 30	154	57	68	57	56
	> 30	24	14	19	12	15
Cancer stage	I	-	26	32	29	21
	II	-	38	46	34	32
	III	-	4	3	3	11
	IV	-	3	6	5	7

Table 2: Identification of women-specific cancer patients from the blinded sample set.

Total #Samples Provided	1000
True Positives Identified	797
True Negatives Identified	200
False Positives Identified	0
False Negatives Identified	3
Sensitivity	99.6 % (797/800)
Specificity	100 % (200/200)

Table 3: Identification of the cancer type (TOO).

Cancer Type	#Samples correctly identified	Accuracy (%)
Breast Cancer	107/107	100
Endometrial Cancer	272/299	90.97
Cervical Cancer	184/192	95.8
Ovarian Cancer	190/199	95.5

Figure 1

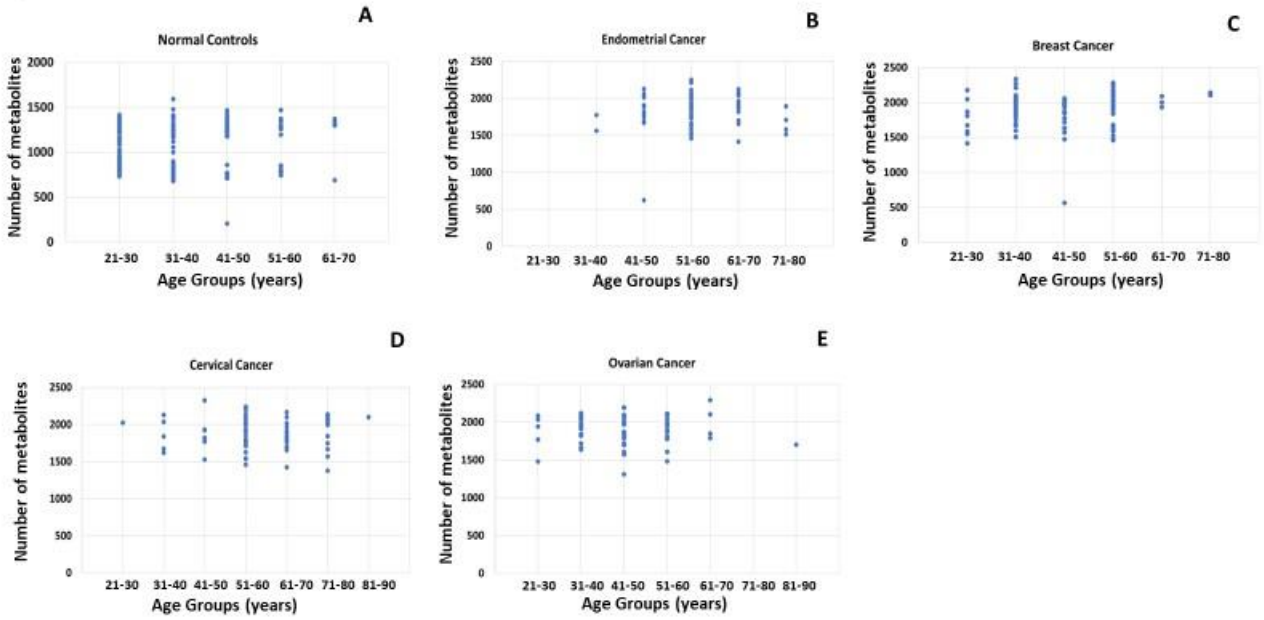

Figure 2

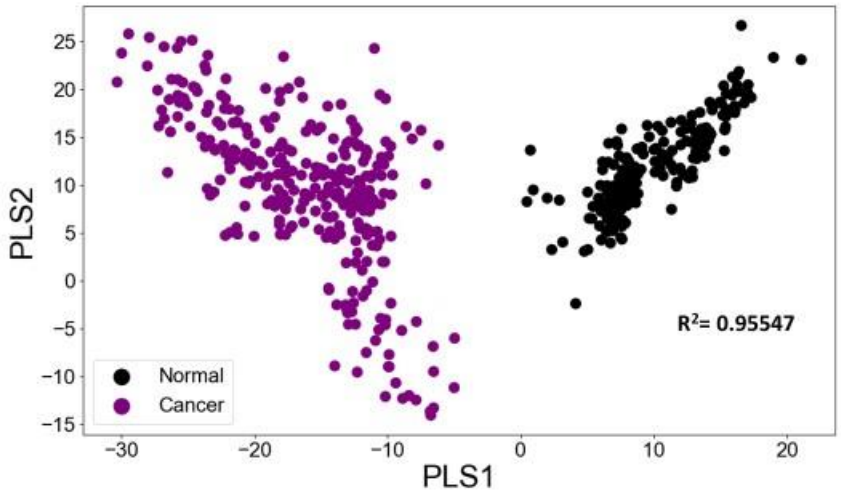

Figure 3

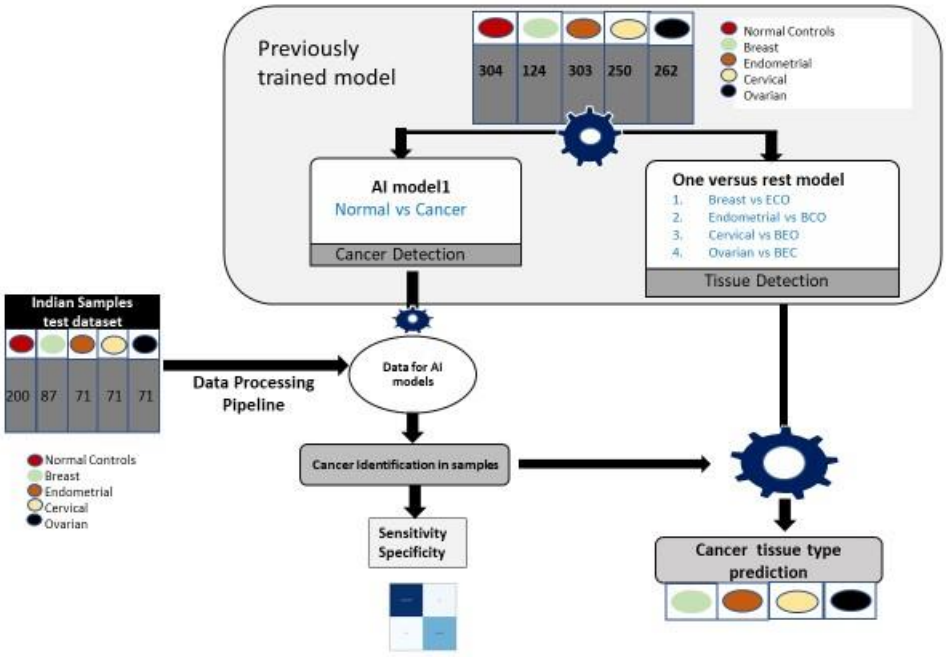

Figure 4

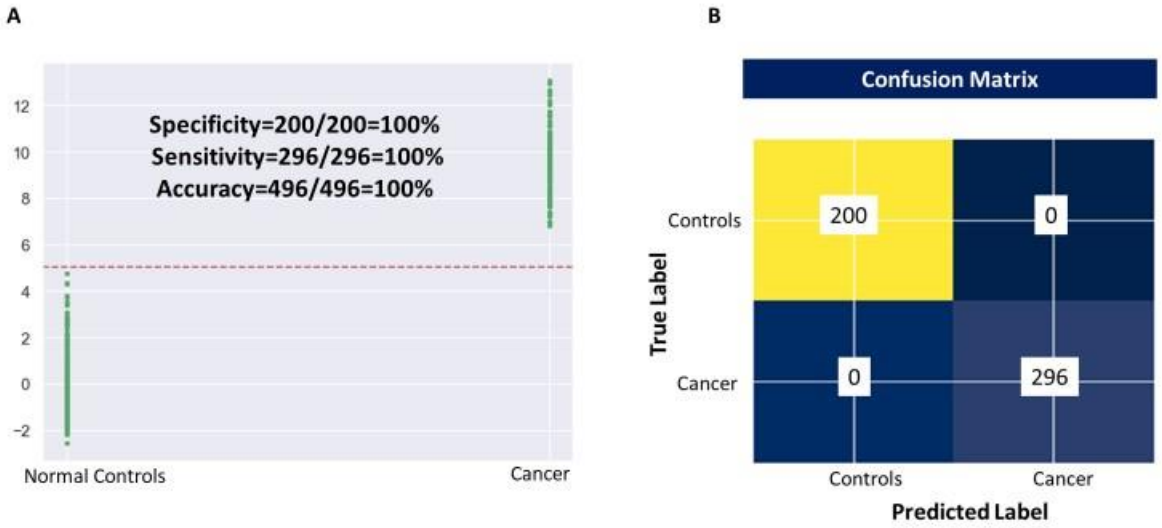

Figure 5

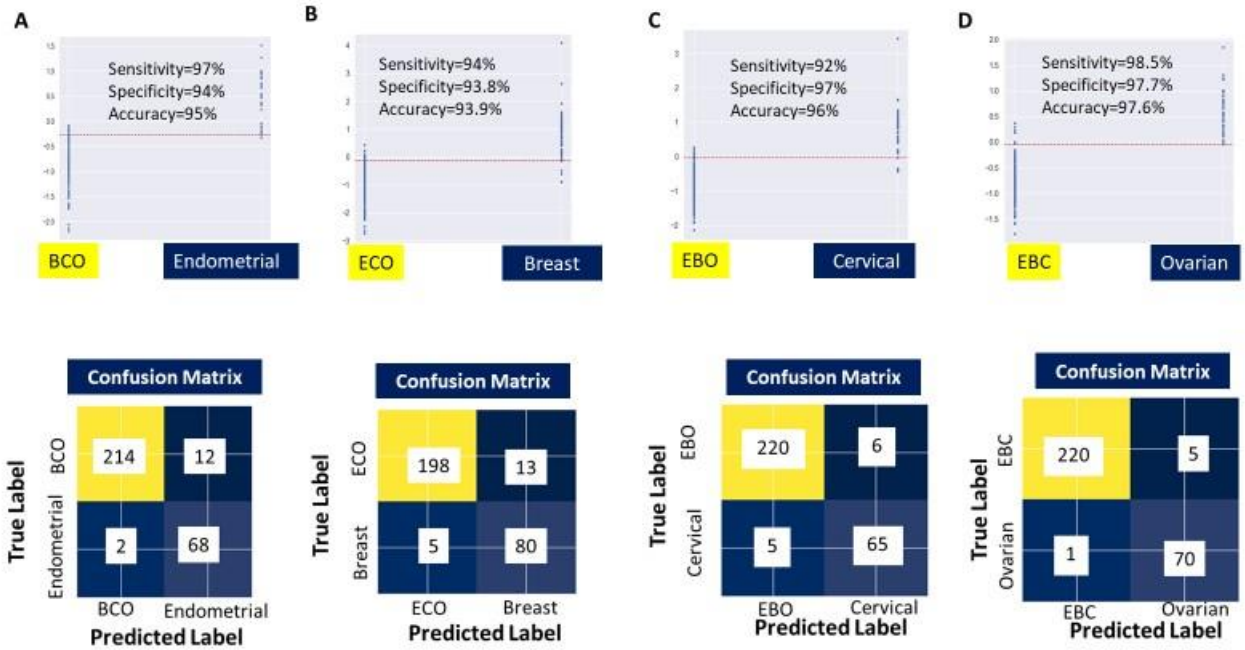

FIGURE LEGENDS

Figure 1. Age-wise detection of detected metabolites.

Figure provides a graphical representation of the number of metabolites detected across the individual age groups, for the normal control set as well as the individual cancer groups. The cumulative unique metabolites detected in normal control samples were 5371. While, endometrial, breast, cervical and ovarian cancer samples were found to have 4947, 5132, 5035 and 5041 respectively.

Figure 2. PLSDA plot distinguishes between the cancer group and also the normal controls.

Figure presents a PLSDA plot of the matrix of sample-specific metabolites versus metabolite intensity for normal controls and the group of women-specific cancer samples. The separation obtained between them is shown. The R^2 value obtained is included.

Figure 3. Schematic of the approach employed for identifying women-specific cancer positive samples in the Indian test dataset.

Illustrated here is the procedure adopted for identifying the cancer-positive samples in the Indian Test dataset. The grey box represents the previously trained model for cancer detection and tissue type identification (24). The Test dataset was first processed using the data processing steps followed previously in the study (24). The final processed data was tested first for the identification of cancer samples in the dataset, the cancer positive identified cases were next further tested using the one versus rest models trained previously for cancer tissue type prediction (24).

Figure 4. Distinguishing women-specific cancers from normal controls.

Results obtained for testing of our previous trained model for distinguishing the women-specific cancer group from normal controls (24) on the data set generated from Indian patient and normal controls is shown here. The separation achieved between the cancer and normal control groups is shown in Panel A and the resulting confusion matrix that was generated is shown in Panel B. Values obtained for Sensitivity, Specificity, and Accuracy are also given.

Figure 5. Testing the multiclass model for its ability to distinguish the individual cancer groups.

Panel (A) shows the results of specifically testing the multiclass trained model for separation of endometrial cancer samples from the other cancers (breast, cervical, ovarian; abbreviated as BCO) based on model's Endometrial scores. The resulting confusion matrix on applying a threshold shows good accuracy, sensitivity, and specificity. Panel (B) shows the results of specifically testing the multiclass trained model for separation of breast cancer samples from the other cancers (endometrial, cervical, and ovarian; abbreviated as ECO) based on model's Breast scores. The resulting confusion matrix on applying a threshold shows good accuracy, sensitivity, and specificity. Panel (C) the results of specifically testing the multiclass trained model for separation of cervical cancer samples from the other cancers (breast, endometrial, ovarian; abbreviated as EBO) based on model's Cervical scores. The resulting confusion matrix on applying a threshold shows good accuracy, sensitivity, and specificity. Panel (D) shows the results of specifically testing the multiclass trained model for separation of ovarian cancer samples from the other cancers (breast, endometrial, cervical; abbreviated as EBC) based on model's Ovarian scores. The resulting confusion matrix on applying a threshold shows high accuracy, sensitivity, and specificity.