BMJ Open

Development of a hoRizontal data intEgration classifier for NOn-invasive early diAgnosis of breasT cancEr: the RENOVATE study protocol

Francesco Ravera,[1] Gabriella Cirmena,[1] Martina Dameri,[2] Maurizio Gallo,[1] Valerio Gaetano Vellone,[3] Piero Fregatti,[2] Daniele Friedman,[2] Massimo Calabrese,[2] Alberto Ballestrero,[1] Alberto Tagliafico,[4] Lorenzo Ferrando,[2] Gabriele Zoppoli [1]

To cite: Ravera F, Cirmena G, Dameri M, *et al*. Development of a hoRizontal data intEgration classifier for NOn-invasive early diAgnosis of breasT cancEr: the RENOVATE study protocol. *BMJ Open* 2021;**11**:e054256. doi:10.1136/bmjopen-2021-054256

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-054256).

AT, LF and GZ contributed equally.

AT, LF and GZ are joint senior authors.

Received 07 June 2021
Accepted 02 November 2021

For numbered affiliations see end of article.

Correspondence to
Professor Gabriele Zoppoli; gabriele.zoppoli@unige.it

ABSTRACT

Introduction Standard procedures aimed at the early diagnosis of breast cancer (BC) present suboptimal accuracy and imply the execution of invasive and sometimes unnecessary tissue biopsies. The assessment of circulating biomarkers for diagnostic purposes, together with radiomics, is of great potential in BC management.

Methods and analysis This is a prospective translational study investigating the accuracy of the combined assessment of multiple circulating analytes together with radiomic variables for early BC diagnosis. Up to 750 patients will be recruited at their presentation at the Diagnostic Senology Unit of Ospedale Policlinico San Martino (Genoa, IT) for the execution of a diagnostic biopsy after the detection of a suspect breast lesion (t0). Each recruited patient will be asked to donate peripheral blood and urine before undergoing breast biopsy. Blood and urine samples will also be collected from a cohort of 100 patients with negative mammography. For cases with histological diagnosis of invasive BC, a second sample of blood and urine will be collected after breast surgery. Circulating tumour DNA, cell-free methylated DNA and circulating proteins will be assessed in samples collected at t0 from patients with stage I–IIA BC at surgery together with those collected from patients with histologically confirmed benign lesions of similar size and from healthy controls with negative mammography. These analyses will be combined with radiomic variables extracted with freeware algorithms applied to cases and matched controls for which digital mammography is available. The overall goal of the present study is to develop a horizontal data integration classifier for the early diagnosis of BC.

Ethics and dissemination This research protocol has been approved by Regione Liguria Ethics Committee (reference number: 2019/75, study ID: 4452). Patients will be required to provide written informed consent. Results will be published in international peer-reviewed scientific journals.

Trial registration number NCT04781062.

Strengths and limitations of this study

- The study has a prospective design with well-balanced controlled cohorts.
- The study assesses the performance of some of the most promising and cutting-edge biomarkers in the field of translational oncology either for diagnostic or predictive purposes in patients affected by breast cancer (BC).
- The performance of the combination of multiple circulating biomarkers and radiomics algorithms for BC early diagnosis is assessed.
- The study is the first to investigate such biomarkers in early (ie, stage I–IIA) BC.
- The study is not designed for the early diagnosis of in situ BCs, which are considered as benign lesions and are included in the control group.

INTRODUCTION

Current protocols for the early diagnosis of breast cancer (BC) rely on the combined use of radiological procedures such as mammography and ultrasound.[1 2] Confirmation biopsy or recall tests are mandatory in case of suspect found during the first examination and bring eventually to a more definite characterisation of the radiologically identified lesion. This approach is however burdened by serious issues which include (a) suboptimal sensitivity and positive predictive power respectively for radiological screening and diagnostic procedures, (b) invasiveness of biopsy with discomfort for women undergoing diagnostic tests, along with the risk of drawing non-representative portions of the suspect region considering the genotypic and phenotypic heterogeneity of BCs,[3] (c) long turnaround time for recall tests, even in high-level centres. In particular, the suboptimal sensitivity of screening procedures leads to non-diagnosed tumours which become able to advance locally and spread systemically, impacting patients' prognosis, while the suboptimal positive predictive power

of diagnostic evaluation implies unnecessary invasive biopsies.[1].

The compelling necessity of increasing the accuracy of screening and diagnostic procedures for cancer has brought to a relevant advancement in the pursuit of accurate biomarkers able to efficiently detect and characterise it. To date, however, no protocol with effective and recognised clinical validity for the early diagnosis of BC has been developed.

Among the most promising biomarkers, circulating tumour DNA (ctDNA), cell-free methylated DNA (cfMeDNA), exosomes and microRNA (miRNA) have been the subject of relevant scientific reports.[4–9] Moreover, machine-learning (so called 'deep learning') algorithms have been applied to traditional radiology imaging techniques for diagnosis assistance with exciting results in the field of radiomics.[10] Results in the proteomics field have instead been few, in part due to the cumbersome methodologies often applied to the study of these molecules.[11]

Since it is hardly conceivable that a single biomarker is able to achieve 100% accuracy in the early detection of BC, the primary aim of the present study is to merge for this purpose the assessment of multiple biological analytes with the refinement of radiomics algorithms, overcoming the aforementioned limitations in terms of accuracy of the individual biomarkers. The concept of combining different data layers to reach a better classifier compared to the individual analytes is referred to as horizontal data integration (HDI) classification. Therefore, the overall goal of the project is to develop an HDI classifier enabling early non-invasive diagnosis of BC with similar accuracy compared with breast biopsies.

METHODS AND ANALYSIS

Study design

This is a prospective translational case–control study with the primary aim of assessing the clinical validity of several biomarkers, individually and combined, in the early detection of BC in a real-life clinical setting (see figure 1 for study diagram). Patients are recruited at the presentation to the Diagnostic Senology Unit of Ospedale Policlinico San Martino (Genoa, Italy) for the execution of the breast biopsy after the radiological detection of a suspect breast lesion ≤2 cm (ie, radiological T1, BIRADS-3/4/5) with no radiological evidence of axillary or distant disease. These patients are asked, on completion of an informed written consent, to donate four peripheral blood tubes (30 mL total) and one urine sample (40 mL). Of these, only samples collected from patients with stage I–IIA neoplasia at surgery (T1N0 or T2N0 or T1N1a) or histologically confirmed benign lesions will be analysed. A second sample is obtained from those patients with confirmed histological diagnosis of invasive BC at the first oncological visit after breast primary surgery (t1) according to normal practice, independently from possible neoadjuvant treatments. No limitations based on tumour stage will be made for the analysis of the latter. Blood and urine samples are also collected and analysed from a cohort of 100 healthy women with two consecutive negative mammograms (BIRADS-1 or BIRADS-0 with negative ultrasound) as well.

All patients are screened with the following inclusion and exclusion criteria. Inclusion criteria include: written informed consent; breast lesions detected by digital bilateral mammography; age ≥18 years and ≤75 years; eligibility for diagnostic biopsy (tru-cut or vacuum assisted breast

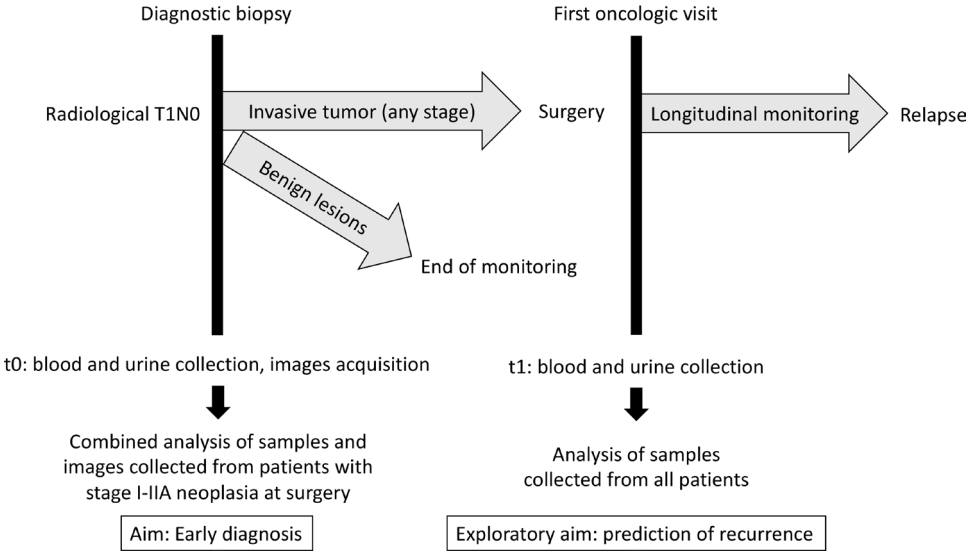

Figure 1 Study diagram. Blood and urine samples will be collected from patients yielding a radiological breast lesion ≤2 cm with no evidence of lymph node neoplastic dissemination (radiological T1N0). Images and samples acquired from patients with stage I–IIA (T1N0 or T2N0 or T1N1a neoplasia) BC at surgery will be analysed for diagnostic purposes together with images and samples acquired from patients yielding benign breast lesions and from patients with negative mammography. Blood and urine samples will be re-collected from all patients yielding invasive neoplasia at diagnosis after surgery at the first oncological visit, and will be analysed for the prediction of breast cancer recurrence.

Ravera F, *et al*. *BMJ Open* 2021;**11**:e054256. doi:10.1136/bmjopen-2021-054256

biopsy) as per normal clinical practice for study population, or absence of breast lesions at the digital mammography for healthy controls (BIRADS-1 or BIRADS-0 with negative ultrasound); ability and willfulness to comply with the protocol requirements.

Exclusion criteria include: history of invasive cancer, any type; clinical or radiological suspicion of advanced or metastatic cancer at the time of screening; known history of active or treated autoimmune or manifest chronic or seasonal and active allergic disorders (with the exception of autoimmune thyroiditis); history of major trauma or surgery during the 24 weeks before screening; history of active infectious disease, either chronic or acute occurring during the 8 weeks before screening; history of known acute or chronic cardiac, kidney or liver disease disorders or acute cardiac events.

Candidate biomarkers

Circulating tumour DNA and cell-free methylated DNA

ctDNA obtained via liquid biopsy has shown serious potential not only in the early diagnosis of cancer but inter alia as an effective marker for its recurrence, longitudinal monitoring and response to therapy.[12] Current methods aimed at detecting ctDNA, however, are mostly based on sequencing somatic mutations from cell-free DNA (cfDNA), a process constrained by relevant limitations in terms of clinical applicability due to (a) expectable low sensitivity in early stage cancers given the limited number of recurrent mutations in ctDNA,[8 13] (b) the vast heterogeneity of DNA mutations occurring in a single tumour together with non-specific mutational profiles along different patients and cancer types,[12 13] (c) current cost-prohibitive impact of cfDNA next generation sequencing (NGS) for mutation assessment.[14] On the other hand, the assessment of cfMeDNA may overcome the limitations outlined above. DNA methylation is commonly involved in cellular development, tissue-specific gene expression, and regulation of imprinted alleles, with widespread effects on cellular growth and genomic stability.[15] Changes in methylomes in cancer are usually associated with alterations of the transcriptional outcome and genomic instability, priming or enhancing carcinogenesis. Given the impact of methylation changes on cellular equilibrium, different methylomes may be associated with specific biological features, providing useful information for the early diagnosis of cancer.[16] Shen *et al* recently developed a sensitive, immunoprecipitation-based protocol to analyse the methylomes of small quantities of cfDNA, providing an efficient method to detect large-scale DNA methylation changes that are enriched for tumour-specific patterns.[8] cfMeDNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) showed high accuracy in the detection of cancer-derived DNA methylation events in cfDNA, managing to distinguish multiple cancer types from healthy controls and discriminate different and specific methylation patterns across diverse cancers.

RNA-based biomarkers

RNA-based biomarkers can be referred either to coding or non-coding RNAs. Non-coding RNAs, including miRNAs and long non-coding RNAs, have been extensively studied over the last years as promising markers for cancer early diagnosis and monitoring, being of particular interest for their stability and quantity in the bloodstream.[17] On the other hand, coding-RNAs include the cell-free messenger RNA, whose evaluation may allow the non-invasive assessment of the whole-body transcriptome[18] with relevant implications for several clinical purposes.

Circulating proteins

Proteomic analyses in neoplastic patients have been evaluated in regard to diverse clinical purposes with suboptimal results, given the challenging characterisation of circulating proteins. The assessment of alterations in protein-mediated signalling networks is of great interest for its possible implications in the early diagnosis and monitoring of cancer, besides the detection of potentially actionable targets for therapeutic purposes.[11]

Sample collection

Peripheral blood and urine collection, processing and storage

One investigator, a research nurse, and two biologists are collaborating in the first 18 months of the project to collect at t0 and t1 two PAXgene Blood ccfDNA Tubes CE-IVD (PreAnalityX, GmbH), one BD Vacutainer K2E (EDTA) Plus Blood Collection Tube CE-IVD (BD life Sciences), one Tempus TM Blood RNA tube (Applied Biosystems) and one urine specimen in a sterile container. Blood samples are processed to extract and store cfDNA from plasma, and to collect proteins, exosomes, peripheral blood mononuclear cells (PBMC), and total RNA. PAXgene tubes are first centrifuged for 15 min at 1900 rcf at room temperature (RT), then the collected plasma is further centrifuged for 10 min at 1900 rcf at RT. EDTA tubes are centrifuged for 15 min at 1600 rcf at RT and the collected plasma is further centrifuged at 1900 rcf for 10 min at RT. Plasma is then aliquoted in cryovials. Tempus tubes for total RNA extraction are immediately stored at −80°C. At the moment of collection, urine is mixed with Cell-Free DNA Urine Preserve (Streck) in order to stabilise cfDNA in samples for up to 7 days at a temperature ranging from 6°C to 37°C. Urine is centrifuged at 2680 rcf for 10 min, and the supernatant is aliquoted into 15 mL tubes and stored at −80°C until cfDNA extraction. Samples are stored at −80°C in a dedicated, Eppendorf CryoCube F740hi ULT Freezer, three compartments.

All test tubes, from blood collection to storage, are barcoded to increase traceability and anonymity to external personnel.

Resources

A 150 m^2 fully equipped laboratory supports the pivotal work on the samples collected for our project. All procedures of platelet separation and plasma processing, cfDNA and RNA extraction, quality control, amplification,

library preparation and sequencing happen in separate, clean environments optimised for such tasks. Every operation, from sample processing, to nucleic acid extraction and sequencing, is semi-automated in order to minimise cross-sample contamination and ensure optimal reproducibility and consistency of operations. A dedicated freezer is available in the laboratories for sample storage. The laboratory is equipped with a ThermoFisher Scientific Ion S5 XL sequencer. Data collection and the dry-lab part of the analyses are performed on high performance workstations maintained in high-security, dedicated environments to avoid the risk of data loss and sensible data cyber-theft.

Analysis and integration of circulating biomarkers with radiomics

Ultrasensitive NGS on ctDNA

We designed a tagged-amplicon NGS panel covering 101 short regions of frequently mutated genes in BC. For this task, we analysed the GENIE dataset V.3.0,[19] which includes 3269 sequenced primary and metastatic BC. This dataset was chosen for its size and homogeneity. Since our purpose was looking at the most frequent, non-overlapping (ie, mutually exclusive) mutations, the sample size of GENIE was deemed as adequate for the panel design. Among the most common cancer genes in BC, very few behave as dominant oncogenes (ie, with frequent mutation in limited regions of the gene, like PIK3CA), whereas several behave as tumour suppressors (ie, with rare mutations occurring along the whole gene, like CDH1). By dividing all the genes sequenced by GENIE into 9603 regions shorter than 50 base pairs, we realised that to detect at least one mutation, we would need to sequence only 101 regions to cover 80% of the whole GENIE dataset, but 420 regions to increase such detection rate to 88%. In particular, amplicons covering TP53 (30 exonic regions), CDH1 (50 exonic regions), GATA3 (13 exonic regions), and PIK3CA (7 exoic regions) would optimise the coverage by minimising the size of our panel. TP53 is planned to be analysed, together with CDH1, for all its exons. On the other hand, PIK3CA presents mostly hotspot mutations in exons 9 and 20. Hence, only those two exons are going to be covered by this design. The 101 selected regions will be assessed in plasma ctDNA using a custom design, which leverages the novel, proprietary tagged-amplicon OncoMine cfDNA methodology by ThermoFisher Scientific. Its use allows for a limit of detection of 0.1% at 20 000× sequencing coverage with 20 ng of circulating DNA, well attainable with two peripheral blood samples collected in PAXgene tubes. ThermoFisher defines ctDNA positivity as the detection of three molecular families. However, this threshold is arbitrary and will be adapted depending on the actual class the samples fall in, in order to optimise the accuracy of the multiomic HDI. cfDNA extraction is performed by using the QIAamp Mini Elute cfDNA Mini or Midi Kit (Qiagen, Hilden, Germany) on the Qiacube system, according to the Manufacturer's instructions. cfDNA is extracted from at least 4 mL of plasma collected in the EDTA tubes and 2 mL of plasma collected in the PAXgene tubes, managing to obtain the requested amount of cfDNA for cfMeDIP-seq (~5 ng) and the assessment of ctDNA mutations (~15 ng). cfDNA is processed on an Ion Chef fluidic handler, and sequenced using the tagged-amplicon methodology with our custom-design panel on a Ion S5 sequencer. Tertiary analysis will be performed on a dedicated Ion Server System available in our laboratory. The library size of our custom panel is half of the commercial ThermoFisher Scientific Ion Torrent OncoMine Comprehensive cfDNA panel, from which the aforementioned performance considerations have been derived. Hence, the use of our custom-design panel appears feasible and should perform better in light of its smaller size and subsequent deeper attainable sequencing coverage.

Methylome profiling of cfDNA

Recent works pointed out encouraging results in terms of accuracy of cfMeDIP-seq in the detection of informative methylation changes of small quantities (1–10 ng) of plasma cfDNA for the diagnosis of renal cell carcinoma and intracranial tumours,[20 21] with the former being effectively detected also by performing cfMeDIP-seq on urinary cfDNA. According to the protocol elaborated by Shen *et al*, cfMeDIP-seq involves four steps: (a) cfDNA end repair, A-tailing and adapter ligation; (b) cfMeDNA immunoprecipitation and enrichment by using an antibody targeting five methylcytosine; (c) library preparation; (d) high throughput NGS on an Illumina platform for cfMeDNA data.[22] This approach based on immunoprecipitation allows to avoid cfDNA bisulfite treatment, typically used to study DNA methylation but associated with a high rate of DNA degradation.

Proteomics analysis

The proposed task is based on the relative quantification of circulating plasma proteins by a novel, highly multiplexed proteomic assay (SomaScan).[23 24] SomaScan (SomaLogic) is considered the most comprehensive protein array available so far for the relative quantification of proteins. Over 7500 proteins can be simultaneously assessed from 55 µL of plasma. This technology involves a new type of aptamers, which are single stranded DNA molecules able to bind proteins, called SOMAmers. In addition to their high affinity for individual proteins, SOMAmers have a unique 40-nucleotide sequence tag and a fluorescent label that allows their identification and quantification in high-density microarrays. SOMAmers have been successfully assembled in a commercial product allowing the comparative evaluation of proteins in a quantity of serum or plasma (or other biological fluids) as low as 55 µL. The SOMAscan assay is a highly multiplexed, sensitive, quantitative and reproducible proteomic tool for biomarker discovery and development. SomaScan can detect proteins within a range of 10 logarithms, allowing for an unmatched sensitivity to detect even femtomolar

protein concentrations. The analysis of the SomaScan is performed by using classic DNA array data analysis and is based on bioinformatics tools that have been developed for gene array analysis.

Radiomics analyses

A preliminary radiomics classifier has been developed by the Diagnostics Senology team of Ospedale Policlinico San Martino, based on digital breast tomosynthesis (DBT) images of consecutive participants from the ASTOUND (Adjunct Screening With Tomosynthesis or Ultrasound in Women With Mammography-Negative Dense Breasts, ClinicalTrials.gov Identifier: NCT02066142) trial.[2] Radiomics analyses were performed on all DBT images within manually selected regions of interest (ROIs) including all the dense parts of the breast and excluding the fatty parts. ROIs were selected by a single radiologist, with proven expertise in quantitative image analysis. Descriptors of the preliminary classifier were selected after initial screening of 104 radiomics features to reduce the risk of over-fitting and according to features previously used to associate breast parenchymal patterns with cancer risk.[25] For the present project, images features will be extracted from the same cases and matched controls for whom NGS and proteomics analyses are performed, using an open source software platform for medical image informatics and advanced deep learning methodologies.

HDI classifier

Sensitivity and specificity of ctDNA analysis, cfMeDNA tests, proteomic analysis, and radiomics for the early detection of BC will be assessed. Data from these classifiers will be subsequently processed and used to generate the HDI model, based on the ensemble learning approach methodology. Ensemble learning combines predictions of multiple individual classifiers obtained by different techniques such as random forest, support vector machine or general linear modelling in order to enhance generalisation power,[26] avoid overfitting, and increase the strength and reliability of the final outcome.[27] Specifically, the outcomes from ctDNA, cfMeDNA, proteomics and radiomics tests will be combined by using a weighted-majority voting approach implemented in the R environment (caret package). We do not envisage RNA-based classifiers to be initially included in our multianalyte model.

Experimental validation and other analyses

On completion of these experiments, we will seek to validate the possible discovery of novel biomarkers and to facilitate their transfer to clinical applicability. Other promising non-invasive biomarkers will be studied, thanks to the unique sample set at our disposal. We foresee the possibility to perform exosome-enriched miRNA sequencing and PBMC transcriptome sequencing by using the Ampliseq Transcriptome solution on our Ion S5 XL sequencer. Eventually, such analyses may be integrated into the HDI classifier to augment the accuracy of BC early diagnosis.

Study outcomes

This study aims at assessing the performance of multiple analytes, individually and combined in a HDI classifier, in the early detection of BC. The primary outcome is defined on the histopathological diagnosis of early invasive BC characterised as per SIAPEC/ASCO/CAP criteria, with radiological extension ≤2 cm (radiological T1) and stage I–IIA at surgery (T1N0 or T2N0 or T1N1a). Benign lesions are defined on the detection of radiological lesions ≤2 cm without the presence of invasive neoplasia at the first biopsy and at surgery, if performed. In situ tumours are included in the latter group.

Clinical data

Electronic case report forms (eCRF) have been designed for the annotation of patients' clinical data. These include study participants' demographics, biometric parameters such as height, weight and body mass index, assumption of alcohol and smoke, information about the endocrine status (premenopause or postmenopause, age at menarche, assumption of endocrine therapy, number of pregnancies), familiarity for BC and predisposing mutations, and presence of comorbidities. Items included in the eCRF are provided in table 1. Tumour histology and immunohistochemical features will be annotated as well.

Data security and confidentiality

Patients' pseudo-anonymised radiological images as well as demographic and anatomopathological data are collected according to the local ethics committee guidelines using a dedicated, state-of-the-art firewalled data collection system, OpenClinica, hosted by University of Genova servers.

Statistical analysis

Projected sample size

The Diagnostics Senology Unit of San Martino Hospital is the highest-level referral centre in Italy for a population basin of more than 2 000 000 people. In a single year, approximately 15 000 mammograms are performed (figure 2). Of these, 1500 yield a radiological suspicion of malignancy with an ensuing biopsy. Assuming a 50% refusal rate by our patients to undergo trial-specific blood collection, we foresee to enrol 750 patients with radiologically suspect small breast lesions. Among these patients, we assume diagnosis of invasive BC will be confirmed in one-third. Of these 250 bioptic diagnoses of BC, which will undergo primary surgery, as per internal historical records, 40% are assumed to be pT1 and 60% higher than pT1. Assuming another 10% failure rate in sample processing and storage, we will have the potential to collect imaging data and samples from approximately 90 patients with pT1 BC and 180 radiological size-matched lesions in a single year (figure 2). Patient enrolment is projected in a time frame of 12 months from the beginning of the study, whereas samples collection at t1 is expected to end approximately 6 months after the enrolment of the last patient.

Table 1 Items investigated at the recruitment

Biometrics	Voluptuous habits	Endocrine status	Predisposition	Comorbidities
Height	Does/did the patient smoke?	Is the patient in the premenopause or postmenopause phase?	Does the patient present predisposing mutations for cancer?	Does the patient present comorbidities?
Weight	If yes, for how many years?	Age at menarche	If yes, which one	If yes, which ones?
BMI	Packets/day	Number of pregnancies	Does the patient have familiarity for breast and/or ovarian cancer?	
	Years since the last cigarette	Does the patient assume substitutive endocrine therapy?		
	How many alcohol units does the patient assume per day?	If yes, for how many years?		
		Does the patient assume endocrine contraceptive therapy?		
		If yes, for how many years?		

Along with items concerning the inclusion and exclusion criteria, information of interest for the assessment of breast cancer risk and possible interference with the evaluation of circulating biomarkers will be annotated.
BMI, body mass index.

Sample size calculation

The sample size required for our analyses is N=147, with 49 biopsy-proven stage I–IIA BC cases and 98 biopsy-proven benign lesions of similar radiological size. In particular, assuming that the best non-HDI classifier is, in the end, the one with the highest number of tested variables (worst-case scenario, for example, transcriptome sequencing on platelets with approximately 20 000 transcripts analysed), we would need[28 29]:

- N=87 samples for a training set, in a 1:2 ratio between histologically proven BCs and benign lesions matched for radiological size, with a standardised fold change of 1.2, n~20 000 features to be assessed, tolerance=0.05 from the best possible classifier, as defined by Dobbin *et al*.[29]
- N=26 samples for a testing set (70/30 split of samples between training and testing set) with similar ratio of cases and controls.
- N=34 samples for a validation set (30% of the sum of training and testing sets).

It may be possible that, adopting HDI classification, the actual needed sample size will be smaller than calculated. At present however, we are not aware of well-established statistical approaches to obtain a more robust estimation of the number needed for our experiments other than the one we adopted.

Amendments

The original protocol for this study underwent three amendments, mainly aimed at including the assessment of urinary biomarkers and expanding the cohort of analysed patients.

Amendment #1: collection and analysis of urine samples

The first amendment to this protocol, presented on 18 February 2020, involved the collection and analysis of urine samples in addition to blood samples, which was not included in the original version. An efficient method for the early diagnosis of BC based on the assessment of urinary biomarkers would substantially eliminate the invasiveness of the overall procedure, possibly facilitating the execution of large-scale screening campaigns.

Amendment #2: healthy controls

The second amendment, presented on 4 February 2021, involved the recruitment and analysis of a cohort of 100 healthy women with negative mammography (BIRADS-1 or BIRADS-0 with adjunct negative ultrasound) as a healthy control group, in addition to patients undergoing breast biopsy. As the presence of benign breast lesions may determine possible variations in blood and urinary analytes, especially in case of in situ tumours, the purpose of the present amendment is to possibly categorise patients on the presence or absence of non-malignant breast lesions, enhancing the accuracy of the HDI classifier.

Amendment #3: redefinition of the study cohort

The third and final amendment, presented on 27 May 2021, involved several points. According to the original protocol, concerning the cohort of patients with histological diagnosis of invasive BC, only samples collected at t0 from patients with T1N0 tumour at surgery would have been analysed. Moreover, the collection and analysis of t1 samples would have involved only this group of

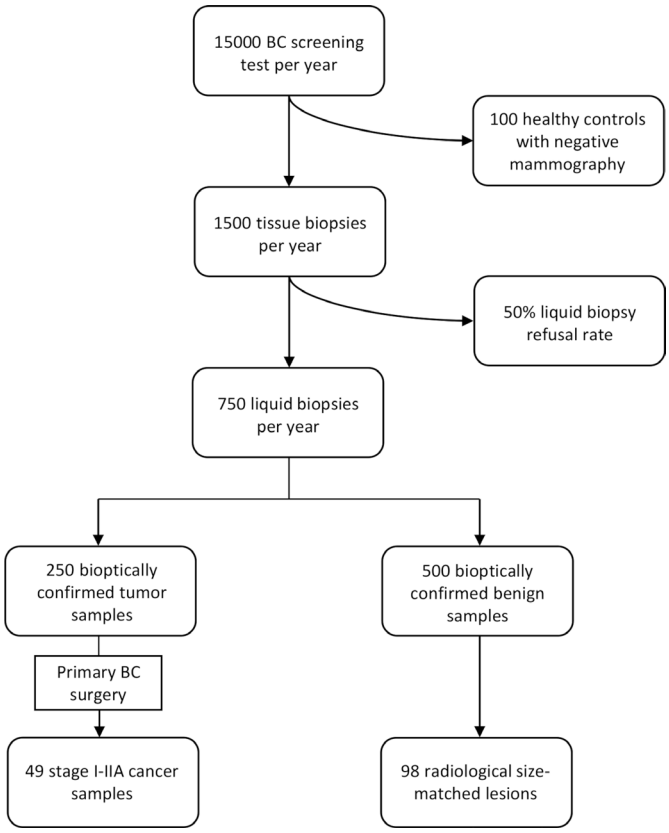

Figure 2 Sample size diagram. Approximately 1500 breast biopsies per year are performed at the Diagnostics Senology Unit of San Martino Hospital. Of a projected number of 750 liquid biopsies, we foresee to collect samples and acquire mammograms from at least 49 patients with stage I–IIA BC, and 98 patients with radiological size-matched lesions, along with those samples and images acquired from 100 healthy women with two consecutive negative mammograms. BC, breast cancer.

patients, with the exclusive aim of distinguishing between tumour-specific and host-specific molecular alterations in connection with the presence/absence of BC. With the present amendment, samples collected at t0 will be analysed from patients assessed as with stage I–IIA tumour at surgery (T1N0 or T2N0 or T1N1a), allowing an effective expansion of the sample size, while remaining in the setting of early stage BC. Furthermore, the collection and analysis of t1 samples will include all patients diagnosed as with invasive BC and sampled at t1, without limitations in terms of tumour stage assessed at surgery. The analysis of t1 samples, coupled with patients' longitudinal monitoring performed as per normal clinical practice, will allow the assessment of the accuracy of a multi-analyte evaluation for the prediction of BC recurrence, added as exploratory aim of the present study.

Study current status

The recruitment phase of the research project started on 18 January 2021. To date (21 October 2021), 183 patients undergoing diagnostic biopsy were recruited. Of these, 98 presented a benign lesion, while 51 presented a malignant lesion. Thirty-four patients were ruled out. Of the recruited patients with histological diagnosis of BC, 35 underwent breast surgery. Of these, 23 patients had a stage I–IIA breast tumour. One hundred and ten patients with negative mammography were recruited as healthy controls. Of these, 20 patients were ruled out. The recruitment phase of the study will end by March 2022. Proteomics, ctDNA and cfMeDIP-seq experiments will be performed during the second year of the research project. The third year of the research project will be dedicated to assess the performance of the individual classifiers, including the one based on radiomics. During the fourth year we will proceed to the wet lab validation, besides performing the transcriptomic experiments. The fifth and final year of the research project will be dedicated to build and optimise the HDI classifier.

Patient and public involvement

There was no patient or public involvement in the design of this study.

ETHICS AND DISSEMINATION

Written informed consents are obtained from each study participant. Participant information sheet includes the main information of the study protocol, the known side effects of peripheral blood collection and the risks implied in the participation to the study, beside the contact information of study investigators. All data are deidentified and no patient-related information will be revealed during analysis. The Regione Liguria Ethics Committee c/o Ospedale Policlinico San Martino has approved the study (reference number: 2019/75, study ID: 4452).

All information concerning patients included in this study are covered by strict confidentiality in compliance with the General Data Protection Regulation EU 2016/679 (GDPR) and D.lgs. 30.06.2003, n. 196, as modified from D.lgs. 10.08.2018, n. 101. The study is conducted in accordance with the national law and according to international guidelines for the conduction of clinical trials referred to as 'Good Clinical Practice'.

Results will be published in international peer-reviewed scientific journals.

DISCUSSION

The assessment of circulating biomarkers in medical oncology is currently burdened by several issues, including its suboptimal accuracy when applied to clinical purposes and the lack of standardised preanalytical and analytical procedures for the evaluation of circulating analytes. Overcoming these limitations would allow the achievement of minimally invasive and personalised assays for the management of neoplastic patients, either in the diagnostic setting or in the early detection of recurrence or the prediction of the response to therapy, possibly replacing or implementing current protocols based on radiology

and traditional tissue biopsies. One of the most effective strategies carried out to enhance the accuracy of liquid biopsies is the contemporary assessment of multiple analytes, which has already shown augmented sensitivity and specificity compared to the evaluation of the individual biomarkers.[30] Such implementation can occur at different levels of complexity, involving different kinds of integration. A basic integration can be referred to the combination of biomarkers of the same kind, such as DNA–DNA or protein–protein combinations. On the other hand, an advanced integration refers to the combined assessment of different kinds of analytes, such as DNA–protein, or the combination of circulating biomarkers with radiological procedures possibly refined with radiomics.[30]

The purpose of this study is to evaluate the performance of the combination of multiple circulating biomarkers, either plasmatic or urinary, and radiomics for BC early diagnosis in patients recruited from a real-life clinical setting. At the enrolment, patients undergo a rigorous selection in order to avoid possible confounding factors that can affect the assessment of tumour-specific circulating DNA, cfMeDNA, circulating RNA and proteins. Only samples collected from patients presenting an early-stage disease (I and IIA) at diagnosis will be analysed for this aim. This selection allows us to assess the effective potential of current techniques for the evaluation of circulating biomarkers in the early diagnosis of the most common cancer worldwide. Conversely, there will be no limits based on tumour stage for the analysis of circulating biomarkers collected after surgery, this being aimed at the prediction of BC recurrence.

This study is not designed for the detection of in situ BCs, which are considered as benign lesions and will be included in the control group. This represents indeed the main limitation of this protocol, given the typical transition of in situ BC to invasive cancer and the consequent necessity of its early detection and eradication.

However, the protocol outlined above presents relevant advantages compared to other studies aimed at the same purpose, including the application of recent and cutting-edge techniques to a selected but realistic cohort of patients, possibly bringing to effective advancements in current standards of BC patients management, besides directing future researches in the panorama of translational medicine applied to oncology.

Author affiliations
[1]Department of Internal Medicine, Università degli Studi di Genova, Genova, Italy
[2]Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia, Genova, Italy
[3]Department of Surgical Sciences and Integrated Diagnostic, Università degli Studi di Genova, Genova, Italy
[4]Department of Health Sciences, Università degli Studi di Genova, Genova, Italy

Contributors GZ conceived the study. FR wrote the manuscript. GC and MD provided a significant contribution in the Methods and analysis section. MG, VGV, PF, DF, MC, AB, AT and LF critically reviewed the manuscript draft.

Funding This trial is entirely supported by a Fondazione AIRC per la Ricerca sul Cancro, Investigator Grant ID 21761 to GZ.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Gabriele Zoppoli http://orcid.org/0000-0002-1619-1708

REFERENCES

1 Sprague BL, Arao RF, Miglioretti DL, *et al*. National performance benchmarks for modern diagnostic digital mammography: update from the breast cancer surveillance Consortium. *Radiology* 2017;283:59–69.
2 Tagliafico AS, Calabrese M, Mariscotti G, *et al*. Adjunct screening with Tomosynthesis or ultrasound in women with Mammography-Negative dense breasts: interim report of a prospective comparative trial. *J Clin Oncol* 2016;34:1882–8.
3 Chung W, Eum HH, Lee H-O, *et al*. Single-Cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* 2017;8:15081.
4 Bettegowda C, Sausen M, Leary RJ, *et al*. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 2014;6:224ra24.
5 Cohen JD, Li L, Wang Y, *et al*. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018;359:926–30.
6 Best MG, Sol N, Kooi I, *et al*. RNA-Seq of Tumor-Educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 2015;28:666–76.
7 Bhome R, Del Vecchio F, Lee G-H, *et al*. Exosomal microRNAs (exomiRs): small molecules with a big role in cancer. *Cancer Lett* 2018;420:228–35.
8 Shen SY, Singhania R, Fehringer G, *et al*. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 2018;563:579–83.
9 Kalluri R. The biology and function of exosomes in cancer. *J Clin Invest* 2016;126:1208–15.
10 Valdora F, Houssami N, Rossi F, *et al*. Rapid review: radiomics and breast cancer. *Breast Cancer Res Treat* 2018;169:217–29.
11 Bhawal R, Oberg AL, Zhang S, *et al*. Challenges and opportunities in clinical applications of blood-based proteomics in cancer. *Cancers* 2020;12. doi:10.3390/cancers12092428. [Epub ahead of print: 27 08 2020].
12 Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: a comprehensive review. *Clin Genet* 2019;95:643–60.
13 Aravanis AM, Lee M, Klausner RD. Next-Generation sequencing of circulating tumor DNA for early cancer detection. *Cell* 2017;168:571–4.
14 Bennett NC, Farah CS. Next-Generation sequencing in clinical oncology: next steps towards clinical validation. *Cancers* 2014;6:2296–312.
15 Stirzaker C, Taberlay PC, Statham AL, *et al*. Mining cancer methylomes: prospects and challenges. *Trends Genet* 2014;30:75–84.
16 Salomon MP, Orozco JIJ, Wilmott JS, *et al*. Brain metastasis DNA methylomes, a novel resource for the identification of biological and clinical features. *Sci Data* 2018;5:180245.
17 Anfossi S, Babayan A, Pantel K, *et al*. Clinical utility of circulating non-coding RNAs - an update. *Nat Rev Clin Oncol* 2018;15:541–63.
18 Koh W, Pan W, Gawad C, *et al*. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc Natl Acad Sci U S A* 2014;111:7361–6.
19 AACR Project GENIE Consortium. AACR project genie: Powering precision medicine through an international Consortium. *Cancer Discov* 2017;7:818–31.

20 Nuzzo PV, Berchuck JE, Korthauer K, *et al*. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat Med* 2020;26:1041–3.
21 Nassiri F, Chakravarthy A, Feng S, *et al*. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat Med* 2020;26:1044–7.
22 Shen SY, Burgener JM, Bratman SV, *et al*. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nat Protoc* 2019;14:2749–80.
23 Webber J, Stone TC, Katilius E, *et al*. Proteomics analysis of cancer exosomes using a novel modified aptamer-based array (SOMAscan™) platform. *Mol Cell Proteomics* 2014;13:1050–64.
24 Raffield LM, Dang H, Pratte KA, *et al*. Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* 2020;20:e1900278.
25 Lambin P, van Stiphout RGPM, Starmans MHW, *et al*. Predicting outcomes in radiation oncology--multifactorial decision support systems. *Nat Rev Clin Oncol* 2013;10:27–40.
26 Dietterich TG. Machine-learning research; four current directions. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.228.6620&rep=rep1&type=pdf
27 Thrun S, Pratt L. Learning to Learn: Introduction and Overview. In: Thrun S, Pratt L, eds. *Learning to learn*. Boston, MA: Springer US, 1998: 3–17.
28 Dobbin KK, Simon RM. Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* 2007;8:101–17.
29 Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res* 2008;14:108–14.
30 Qiu J, Xu J, Zhang K, *et al*. Refining cancer management using integrated liquid biopsy. *Theranostics* 2020;10:2374–84.