

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Development of a horizontal data integration classifier for NONinvasive early diagnosis of breast cancer: the RENOVATE study protocol

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-054256
Article Type:	Protocol
Date Submitted by the Author:	07-Jun-2021
Complete List of Authors:	<p>Ravera, Francesco; Università degli Studi di Genova, Department of internal Medicine</p> <p>Cirmena, Gabriella; Università degli Studi di Genova, Department of internal Medicine</p> <p>Dameri, Martina; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia</p> <p>Gallo, Maurizio; Università degli Studi di Genova, Department of internal Medicine</p> <p>Vellone, Valerio; Università degli Studi di Genova, Department of Surgical Sciences and Integrated Diagnostic</p> <p>Fregatti, Piero; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia</p> <p>Friedman, Daniele; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia</p> <p>Calabrese, Massimo; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia</p> <p>Ballestrero, Alberto; University of Genoa, Department of Internal Medicine</p> <p>Tagliafico, Alberto; Università degli Studi di Genova, Department of Health Sciences</p> <p>Ferrando, Lorenzo; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia</p> <p>Zoppoli, Gabriele; Università degli Studi di Genova, Department of internal Medicine</p>
Keywords:	Breast tumours < ONCOLOGY, Breast imaging < RADIOLOGY & IMAGING, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Cancer genetics < GENETICS, Adult oncology < ONCOLOGY

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3 **Development of a hoRizontal data intEgration classifier for NOninvasiVe early**
4 **diAgnosis of breasT cancEr: the RENOvATE study protocol**
5
6
7
8

9 Authors: Francesco Ravera¹, Gabriella Cirmena¹, Martina Dameri², Maurizio Gallo²,
10 Valerio Gaetano Vellone^{2,4}, Piero Fregatti^{2,4}, Daniele Friedman^{2,4}, Massimo Calabrese²,
11 Alberto Ballestrero^{1,2} Alberto Tagliafico^{2,3,*}, Lorenzo Ferrando^{2,*}, Gabriele Zoppoli^{1,2*,#}
12
13
14
15
16
17
18
19

20 ¹Department of Internal Medicine and Medical Specialties DiMI, University of Genoa,
21 16132, Italy
22
23

24 ²IRCCS Ospedale Policlinico San Martino, 16132 Genoa, Italy
25

26 ³Department of Health Sciences DISSAL, University of Genoa, 16132, Italy
27

28 ⁴Department of Surgical Sciences and Integrated Diagnostic DISC, 16132 Genoa, Italy
29
30
31
32

33 *These Authors equally contributed to the present manuscript
34

35 # Correspondence should be addressed to Gabriele Zoppoli: gabriele.zoppoli@unige.it
36
37
38
39
40

41 **Keywords:** breast cancer, early diagnosis, liquid biopsy, circulating biomarkers, radiomics
42

43 **Word count:** 4459
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Abstract

Introduction. Current standards of care for breast cancer (BC) early diagnosis 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. 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. Patients are 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 is asked to donate peripheral blood and urine before undergoing biopsy. Blood and urine samples are also collected from a cohort of patients with negative mammography. Only for cases with histological diagnosis of invasive BC, a second sample of blood and urine is collected at the first visit after breast surgery. Circulating tumor DNA, cell-free methylated DNA, RNA from peripheral blood mononuclear cells 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 radiomics 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), and is registered in ClinicalTrials.gov with the ID NCT04781062. Results will be published in international peer-reviewed scientific journals.

Strengths and limitations

Strengths

- 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 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 (i.e. stage I-IIA) BC

Limitations

- 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

1. Introduction

Standard of care early diagnosis of breast cancer (BC) relies 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 characterization 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 approaches, (b) invasiveness of biopsy with discomfort for women undergoing diagnostic tests, along with the risk of drawing non-representative portions of the suspected region considering the genotypic and phenotypic heterogeneity of BCs [3], (c) long turnaround time for recall tests, even in high-level centers. In particular, the suboptimal sensitivity of screening procedures leads to non-diagnosed tumors 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].

1
2
3 The compelling necessity of increasing the accuracy of screening and diagnostic procedures
4 for cancer has brought to a relevant advancement in the pursuit of accurate biomarkers able
5 to efficiently detect and characterize it, in the panorama of liquid biopsies. To date, however,
6 no protocol with effective and recognized clinical validity for the early diagnosis of BC has
7 been developed.
8
9
10
11
12

13 Amongst the most promising biomarkers, circulating tumor DNA (ctDNA) and cell-free
14 methylated DNA (cfMeDNA), exosomes, and microRNA (miRNA) have been the subject of
15 relevant scientific reports [4–9]. Moreover, machine-learning (so called “deep learning”)
16 algorithms have been applied to traditional radiology imaging techniques for diagnosis
17 assistance with exciting results in the field of radiomics [10]. Results in the proteomics field
18 have instead been few, in part due to the cumbersome methodologies often applied to the
19 study of these molecules [11].
20
21
22
23
24
25
26
27

28 Since it is hardly conceivable that a single biomarker is able to achieve 100% accuracy in the
29 early detection of BC, the primary aim of the present study is to merge for this purpose the
30 assessment of multiple biological analytes with the refinement of radiomics algorithms,
31 overcoming the aforementioned limitations in terms of accuracy of the individual biomarkers.
32 The concept of combining different data layers to reach a better classifier compared to the
33 individual analytes is referred to as horizontal data integration (HDI) classification. Therefore,
34 the overall goal of the project is to develop an HDI classifier enabling early noninvasive
35 diagnosis of BC with similar accuracy compared to breast biopsies.
36
37
38
39
40
41
42
43
44
45
46
47

48 **2. Methods**

49 **2.1. Study design**

50 This is a prospective translational case-control study with the primary aim of assessing the
51 clinical validity of several biomarkers, individually and combined, in the early detection of BC
52 in a real-life clinical setting (see Figure 1 for study diagram). Patients are recruited at the
53 presentation to the Diagnostic Senology Unit of Ospedale Policlinico San Martino for the
54 execution of the breast biopsy after the radiological detection of a suspected breast lesion ≤ 2
55
56
57
58
59
60

1
2
3 cm (i.e. radiological T1) with no radiological evidence of axillary or distant disease. Blood and
4 urine samples are collected at this time point (t0) from each recruited patient who fulfills the
5 inclusion and exclusion criteria mentioned in paragraph 2.3. Of these, only samples collected
6 from patients with stage I-IIA neoplasia at surgery (T1N0 or T2N0 or T1N1a) or histologically
7 confirmed benign lesions will be analysed. A second sample is obtained from those patients
8 with confirmed histological diagnosis of invasive BC at the first oncologic visit after breast
9 primary surgery (t1) according to normal practice, independently from possible neoadjuvant
10 treatments. No limitations based on tumor stage will be made for the analysis of the latter.
11 Blood and urine samples are also collected and analysed from a cohort of 100 healthy women
12 with two consecutive negative mammograms (BIRADS-1 or BIRADS-0 with negative
13 ecography) as well.

28 **2.2. Candidate biomarkers**

30 **2.2.1. Circulating tumor DNA and cell-free Methylated DNA**

31 ctDNA obtained via liquid biopsy has shown serious potential not only in the early diagnosis
32 of cancer but *inter alia* as an effective marker for its recurrence, longitudinal monitoring and
33 response to therapy [12]. Current methods for detecting ctDNA however are mostly based on
34 sequencing somatic mutations from cell-free DNA (cfDNA), a process constrained by relevant
35 limitations in clinical applicability due to (a) expectable low sensitivity in early stage cancers
36 given the limited number of recurrent mutations in ctDNA [8,13], (b) the vast heterogeneity of
37 DNA mutations occurring in a single tumor together with nonspecific mutations profiles along
38 different patients and cancer types [12,13], (c) current cost-prohibitive impact of cfDNA Next
39 Generation Sequencing (NGS) for mutation assessment [14]. On the other hand, the
40 assessment of cfMeDNA may overcome the limitations outlined above. DNA methylation is
41 commonly involved in cellular development, tissue-specific gene expression, and regulation
42 of imprinted alleles, with widespread effects on cellular growth and genomic stability [15].
43 Changes in methylomes in cancer are usually associated with alterations of the transcriptional
44 outcome and genomic instability, priming or enhancing carcinogenesis. Given the impact of
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 methylation changes on cellular equilibrium, different methylomes may be associated with
4 specific biological features providing useful information for the early diagnosis of cancer [16].
5 Shen *et al.* recently developed a sensitive, immunoprecipitation-based protocol to analyze the
6 methylomes of small quantities of cfDNA, providing an efficient method to detect large-scale
7 DNA methylation changes that are enriched for tumour-specific patterns [8]. cfMeDNA
8 immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) was able to find cancer-
9 derived DNA methylation events in cfDNA and to distinguish multiple cancer types from
10 healthy controls, besides effectively discriminating different methylation patterns across
11 diverse cancers.

22 **2.2.2. RNA-based biomarkers**

23
24 RNA-based biomarkers can refer either to coding or non-coding RNAs. Non-coding RNAs,
25 including miRNAs and long non coding RNAs, have been extensively studied over the last
26 years as promising markers for cancer early diagnosis and monitoring, being of particular
27 interest for their stability and quantity in the bloodstream [17]. On the other hand, coding-RNAs
28 include the cell-free messenger RNA, whose evaluation may allow the non-invasive
29 assessment of the whole-body transcriptome [18] with relevant implications for several clinical
30 purposes.

38 **2.2.3. Circulating proteins**

39
40 Proteomic analyses in neoplastic patients have been applied to diverse clinical aims with
41 suboptimal results, given the challenging characterization of circulating proteins. The
42 assessment of alterations in protein-mediated signaling networks is of great interest for its
43 possible implications in the early diagnosis and monitoring of cancer, besides the detection of
44 potentially actionable targets for therapeutic purposes [11].

51 **2.3. Patient recruitment and sampling**

52
53 To conform as much as possible to a real-life clinical setting, patients are recruited from the
54 Diagnostics Senology Unit of Ospedale Policlinico San Martino, one of the highest-level
55 referral centers in Italy. Eligible participants include women with radiologically identified
56 lesions, BIRADS-3/4/5, ≤ 2 cm by radiological assessment (i.e., radiological T1), evaluated at
57
58
59
60

1
2
3 the Diagnostic Senology Unit of this Institution. These patients are asked, upon completion of
4 an informed written consent, to donate four peripheral blood tubes (30 mL total) and one urine
5 sample (40 mL). Their pseudo-anonymized radiological images as well as demographic and
6 anatomopathological data are collected as well, by using a freeware dedicated, state-of-the-
7 art firewalled data collection system, OpenClinica®, hosted by University of Genoa servers.
8
9

10
11 All patients are screened with the following inclusion and exclusion criteria. Inclusion criteria
12 include: written informed consent; breast lesions detected by digital bilateral mammography;
13 age \geq 18 years and \leq 75 years; eligibility for diagnostic biopsy (tru-cut or Vacuum Assisted
14 Breast Biopsy) as per normal clinical practice for study population, or absence of breast
15 lesions at the digital mammography for healthy controls (BIRADS-1 or BIRADS-0 with negative
16 ecography); ability and willfulness to comply with the protocol requirements.
17
18

19
20 Exclusion criteria include: previous history of invasive cancer, any type; clinical or radiological
21 suspicion of advanced or metastatic cancer at the time of screening; known history of active
22 or treated autoimmune or manifest chronic or seasonal and active allergic disorders (with the
23 exception of autoimmune thyroiditis); history of major trauma or surgery during the 24 weeks
24 before screening; history of active infectious disease, either chronic or acute occurring during
25 the 8 weeks before screening; history of known acute or chronic cardiac, kidney, or liver
26 disease disorders or acute cardiac events.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 **2.4. Sample collection**

44 **2.4.1. Peripheral blood and urine collection, processing and storage**

45
46 One investigator, a research nurse, and two biologists are collaborating in the first eighteen
47 months of the project to collect at t0 and t1 two PAXgene® Blood ccfDNA Tubes CE-IVD (
48 PreAnalytix, GmbH), one BD Vacutainer® K2E (EDTA) Plus Blood Collection Tube CE-IVD
49 (BD life Sciences), one Tempus TM Blood RNA tube (Applied Biosystems), and one urine
50 specimen in a sterile container. Blood samples are processed to extract and store cfDNA from
51 plasma, and to collect proteins, exosomes, PBMCs, and total RNA . During sample processing
52 PAXgene tubes are first centrifuged for 15 minutes at 1900 rpm at room temperature (RT)
53
54
55
56
57
58
59
60

1
2
3 then the collected plasma is centrifuged for further 10 minutes at 1900 rpm at RT. EDTA tubes
4 are centrifuged for 15 minutes at 1600 rpm at RT and the collected plasma is centrifuged at
5 1900 rpm for further 10 minutes at RT. Plasma is then aliquoted in cryovials. Tempus tubes
6 for total RNA extraction are immediately stored at -80°C. At the moment of collection, urine is
7 mixed with Cell-Free DNA Urine Preserve (Streck) in order to stabilize cfDNA in samples for
8 up to 7 days at 6 °C to 37 °C. Urine is centrifuged at 2680 rpm for 10 minutes, and the
9 supernatant is aliquoted into 15 mL tubes and stored at -80°C until cfDNA extraction. Samples
10 are stored at -80°C in a dedicated, Eppendorf CryoCube F740hi ULT Freezer, 3
11 Compartments.
12

13 All test tubes, from blood collection to storage, are barcoded to increase traceability and
14 anonymity to external personnel.
15
16

17 **2.4.2. Resources**

18 A 150 sqm fully equipped laboratory supports the pivotal work on the samples collected for
19 our project. All procedures of platelet separation and plasma processing, cfDNA and RNA
20 extraction, quality control, amplification, library preparation and sequencing happen in
21 separate, clean environments optimized for such tasks. Every operation from sample
22 processing, to nucleic acid extraction and sequencing is semi-automated in order to minimize
23 cross-sample contamination issues and ensure optimal reproducibility and consistency of
24 operations. A dedicated freezer is available in the laboratories for sample storage. The
25 laboratory is equipped with a ThermoFisher Scientific Ion™ S5 XL sequencer. Data collection
26 and the dry-lab part of the analyses are performed on high performance workstations
27 maintained in high-security, dedicated environments to avoid the risk of data loss and sensible
28 data cyber-theft.
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55

56 **2.5. Analysis and integration of circulating biomarkers with radiomics**

57 **2.5.1. Ultrasensitive NGS on ctDNA**

58
59
60

1
2
3 We designed a tagged-amplicon NGS panel covering 101 short regions of frequently mutated
4 genes in BC. For this task, we analyzed the GENIE dataset v3.0 [19], which includes 3,269
5 sequenced primary and metastatic BC. Amongst the more common cancer genes in BC, very
6 few behave as dominant oncogenes (i.e. with frequent mutation in limited regions of the gene,
7 like PIK3CA), whereas several behave as tumor suppressors (i.e., with rare mutations
8 occurring along the whole gene, like CDH1). By dividing all the genes sequenced by GENIE
9 into 9,603 regions shorter than 50 base pairs, we realized that to detect at least one mutation,
10 we would need to sequence only 101 regions to cover 80% of the whole GENIE dataset, but
11 420 regions to increase such detection rate to 88%. In particular, amplicons covering TP53
12 (30 regions), CDH1 (50 regions), GATA3 (13 regions), and PIK3CA (7 regions) would optimize
13 the coverage by minimizing the size of our panel. The 101 selected regions will be assessed
14 in plasma ctDNA using a custom design, which leverages the novel, proprietary tagged-
15 amplicon OncoMine™cfDNA methodology by ThermoFisher Scientific Inc. Its use allows for a
16 limit of detection of 0.1% at 20,000x sequencing coverage with 20 ng of circulating DNA, well
17 attainable with two peripheral blood samples collected in PAXgene tubes. cfDNA extraction is
18 performed by using the QIAamp Mini Elute cfDNA Mini or Midi Kit (Qiagen, Hilden, Germany)
19 on the Qiacube system, according to the Manufacturer's instructions. cfDNA is processed on
20 an Ion Chef™ fluidic handler, and sequenced using the tagged-amplicon methodology with our
21 custom-design panel using as per manufacturer's specifications on a Ion S5™ sequencer.
22 Tertiary analysis will be performed on a dedicated Ion Server System available in our
23 laboratory. The library size of our custom panel is half of the commercial ThermoFisher
24 Scientific Ion Torrent™ OncoMine™ Comprehensive cfDNA™ panel, from which the
25 aforementioned performance considerations have been derived. Hence, the use of our
26 custom-design panel appears feasible and should perform better in light of its smaller size and
27 subsequent deeper attainable sequencing coverage.

28 **2.5.2. Methylome profiling of cfDNA**

1
2
3 Recent works pointed out astonishing results in terms of accuracy of cfMeDIP-seq in the
4 detection of informative methylation changes of small quantities (1-10 ng) of plasma cfDNA
5 for the diagnosis of renal cell carcinoma and intracranial tumors [20,21], with the former being
6 effectively detected also by cfMeDIP-seq applied to urinary cfDNA. For this study cfDNA will
7 be extracted with the QIAamp Mini Elute cfDNA Mini or Midi Kit (Qiagen, Hilden, Germany) on
8 the Qiacube system from plasma and urine. According to the protocol elaborated by Shen *et*
9 *al.*, cfMeDIP involves four steps: 1) cfDNA end repair, A-tailing and adapter ligation; 2)
10 cfMeDIP immunoprecipitation and enrichment using an antibody targeting 5 methylcytosine;
11 3) library preparation; 4) high throughput NGS on an Illumina platform for cfMeDNA data [22].
12 This approach based on immunoprecipitation allows to avoid cfDNA bisulfite treatment,
13 typically used to study DNA methylation but burdened by the high rate of DNA degradation
14 during the process.

2.5.3. Proteomics analysis

15 The proposed task is based on the relative quantification of blood proteins by a novel, highly
16 multiplexed proteomic assay (SomaScan™) [23,24]. SomaScan™ (SomaLogic™) is
17 considered the most comprehensive protein array available so far for the relative quantification
18 of proteins. Over 7,000 proteins can be simultaneously analyzed starting from 55 µL of plasma.
19 This technology involves a new type of aptamers, which are single stranded DNA molecules
20 able to bind proteins, called SOMAmers. In addition to their high affinity for individual proteins,
21 SOMAmers have a unique 40-nucleotide sequence tag and a fluorescent label that allows
22 their identification and quantification in high-density microarrays. SOMAmers have been
23 successfully assembled in a commercial product allowing the comparative evaluation of
24 proteins in as low as 55µl of serum or plasma (or other biological fluids). The SOMAscan™
25 assay is a highly multiplexed, sensitive, quantitative, and reproducible proteomic tool for
26 biomarker discovery and development. SomaScan™ can detect proteins within a range of 10
27 logarithms, allowing for an unmatched sensitivity to detect even femtomolar protein
28 concentrations. The analysis of the SomaScan™ is performed using classic DNA array data

1
2
3 analysis and is based on bioinformatics tools that have been developed for gene array
4 analysis.
5

6 7 **2.5.4. Radiomics analyses**

8
9 A preliminary radiomics classifier has been developed by the Diagnostics Senology team of
10 Ospedale Policlinico San Martino, based on digital breast tomosynthesis (DBT) images of
11 consecutive participants from the ASTOUND trial [2]. Radiomics analyses were performed on
12 all DBT images within manually selected regions of interest (ROIs) including all the dense
13 parts of the breast and excluding the fatty parts. ROIs were selected by a single radiologist,
14 with proven expertise in quantitative image analysis. Descriptors of the preliminary classifier
15 were selected after initial screening of 104 radiomics features to reduce the risk of over-fitting
16 and according to features previously used to associate breast parenchymal patterns with
17 cancer risk [25]. For the present project, image features will be extracted from the same cases
18 and matched controls for whom NGS and proteomics analyses are performed, using an open
19 source software platform for medical image informatics, and will make use of the advanced
20 deep learning methodologies mastered by our External Collaborators (An.B.).
21
22
23
24
25
26
27
28
29
30
31
32
33
34

35 **2.5.5. HDI classifier**

36 Data from ctDNA analysis, cfMeDNA tests, proteomics, and radiomics classifiers will be
37 processed and used to generate the HDI model, based on the ensemble learning approach
38 methodology. Ensemble learning combines predictions of multiple individual classifiers
39 obtained by different techniques such as random forest, support vector machine or general
40 linear modeling in order to enhance generalization power [26], avoid overfitting, and increase
41 the strength and reliability of the final outcome [27]. Specifically, the outcomes from ctDNA,
42 cfMeDNA, proteomics and radiomics tests will be combined by using a weighted-majority
43 voting approach implemented in the R environment (caret package).
44
45
46
47
48
49
50
51
52
53

54 **2.5.6. Experimental validation and other analyses**

55 Upon completion of proteomics experiments, we will seek to validate the possible discovery
56 of novel protein biomarkers and to facilitate their transfer to clinical applicability by performing
57 conventional protein immunoassays. Other promising noninvasive biomarkers will be studied,
58
59
60

1
2
3 thanks to the unique sample set at our disposal. We foresee the possibility to perform
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

thanks to the unique sample set at our disposal. We foresee the possibility to perform
exosome-enriched miRNA sequencing, 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 reach an even higher accuracy for early BC
diagnosis.

2.6. Study outcomes

This study aims to assess the performance of multiple analytes either individually or combined
in a HDI classifier in the early detection of BC. The primary outcome is defined upon the
histopathological diagnosis of early invasive BC characterized 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 upon the detection of radiological lesions ≤ 2
cm without the presence of invasive neoplasia at the first biopsy and at surgery, if performed.
In situ tumors are included in the latter group.

2.7. 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 BMI, assumption of alcohol and smoke, information about the endocrine status
(pre- or post-menopause, age at menarche, assumption of endocrine therapy, number of
pregnancies), familiarity for BC and predisposing mutations, comorbidities. Items included in
the eCRF are provided in **Tab. 1**

2.8. Data security and confidentiality

Patients' pseudo-anonymized 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.

3. STATISTICAL ANALYSIS

3.1. Projected sample size

The Diagnostics Senology Unit of San Martino Hospital is the highest-level referral center 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, 1,500 yield a radiologic suspicion of malignancy with an ensuing biopsy. Assuming a 50% refusal rate by our patients to undergo trial-specific blood collection, we foresee to enroll 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 enrollment and t0 sample collection is projected in a time frame of 12 months from the beginning of the study, whereas t1 sample collection is expected approximately 6 months after the enrollment of the last patient.

3.2. 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, e.g., transcriptome sequencing on platelets with approximately 20,000 transcripts analyzed), we would need [28,29]:

- $N = 87$ samples for a training set, in a 1:2 ratio between bioptically proven BCs and

1
2
3 radiological size-matched benign lesions, with a standardized fold change of 1.2, $n \sim 20,000$
4 features to be assessed, tolerance = 0.05 from the best possible classifier, as defined by
5 Dobbin *et al.* [29];
6
7

8
9 - N = 26 samples for a testing set (70/30 split of samples between training and testing set)
10 with similar ratio of cases and controls;
11

12
13 - N = 34 samples for a validation set (30% of the sum of training and testing sets).
14

15
16 It may be possible that, adopting HDI classification, the actual needed sample size will be
17 smaller than calculated. At present however, we are not aware of well-established statistical
18 approaches to obtain a more robust estimation of the number needed for our experiments
19 other than the one we adopted.
20
21
22
23
24
25

26 **4. Ethics and dissemination**

27
28 Written informed consents are obtained from each study participant. Participant information
29 sheet includes the main information of the study protocol, the known side effects or any risks
30 implied in the participation to the study, and the contact information of study investigators. All
31 data are deidentified and no patient-related information will be revealed during analysis. The
32 Regione Liguria Ethics Committee c/o Ospedale Policlinico San Martino has approved the
33 study (reference number: 2019/75, study ID: 4452).
34
35
36

37
38 All information concerning patients included in this study are covered by strict confidentiality
39 in compliance with the General Data Protection Regulation EU 2016/679 (GDPR) and D.lgs.
40 30.06.2003, n. 196, as modified from D.lgs. 10.08.2018, n. 101. The study is conducted in
41 accordance with the national law and according to international guidelines for the conduction
42 of clinical trials referred to as "Good Clinical Practice".
43
44
45
46
47
48
49
50
51

52 **5. Amendments**

53
54 The original protocol for this study underwent three amendments, mainly aimed at including
55 the assessment of urinary biomarkers and expanding the cohort of analysed patients.
56
57
58
59
60

5.1. Amendment #1: collection and analysis of urine samples

The first amendment to this protocol, presented on February 18th 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.

5.2. Amendment #2: healthy controls

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

5.3. Amendment #3: redefinition of the study cohort

The third and final amendment, presented on May 27th 2021, involved several substantial 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 tumor at surgery would have been analysed. Moreover, the collection and analysis of t1 samples would have involved only this group of patients, with the exclusive aim of distinguishing between tumor- 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 tumor at surgery (T1N0 or T2N0 or T1N1a), allowing an effective expansion of the sample size, while remaining in the setting of early stage BC diagnosis. 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 tumor stage assessed at surgery. The

1
2
3 analysis of t1 samples, coupled with patients' longitudinal monitoring performed as per normal
4 clinical practice, will allow the assessment of the accuracy of a multi-analyte evaluation for the
5 prediction of BC recurrence, added as exploratory aim of the present study.
6
7
8
9

10 11 **6.DISCUSSION**

12
13 The assessment of circulating biomarkers for clinical purposes is currently burdened by
14 several issues, including suboptimal accuracy and lacking standardization of pre-analytical
15 and analytical procedures. The overcoming of these limitations would allow the achievement
16 of minimally invasive and personalized assays for the management of neoplastic patients,
17 either in the diagnostic setting or in the early detection of recurrence or the prediction of the
18 response to therapy, possibly replacing or implementing current protocols based on traditional
19 tissue biopsies. One of the most effective strategies carried out to enhance the accuracy of
20 liquid biopsies is the contemporary assessment of multiple analytes, which has already shown
21 augmented sensitivity and specificity compared to the evaluation of single biomarkers [30].
22 Such implementation can occur at different levels of complexity, involving different kinds of
23 integration. A basic integration can be referred to the combination of biomarkers of the same
24 kind, such as DNA-DNA or protein-protein combinations. On the other hand, an advanced
25 integration refers to the combined assessment of different kinds of analytes, such as DNA-
26 protein, or the combination of circulating biomarkers with radiological procedures possibly
27 refined with radiomics [30].
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 The purpose of this study is to evaluate the performance of the combination of multiple
46 circulating biomarkers, either plasmatic or urinary, and radiomics for BC early diagnosis in
47 patients recruited from a real-life clinical setting. At the enrollment, patients undergo a rigorous
48 selection in order to avoid possible confounding factors that can affect the assessment of
49 tumor-specific cfMeDNA, circulating RNA, and proteins. Only samples collected from patients
50 presenting an early-stage disease (I and IIA) at diagnosis will be analysed for this aim. This
51 selection allows us to assess the effective potential of current techniques for the evaluation of
52 circulating biomarkers in the early diagnosis of the most common cancer worldwide.
53
54
55
56
57
58
59
60

1
2
3 Conversely, there will be no limits based on tumor stage for the analysis of circulating
4 biomarkers collected after surgery, this being aimed at the prediction of BC recurrence.
5

6
7 In conclusion, the protocol outlined above presents relevant advantages compared to other
8 studies aimed at the same purpose, including the application of recent and cutting-edge
9 techniques to a selected but realistic cohort of patients, possibly bringing to effective
10 advancements to current protocols for the management of BC patients, and directing future
11 researches in the panorama of translational medicine applied to oncology.
12
13
14
15
16

17
18
19
20 **Author statement.** GZ conceived the study. FR wrote the manuscript. GC and MD provided
21 a significant contribution in the Methods section of the manuscript. MG, VGV, PF, DF, MC,
22 AB, AT, and LF critically reviewed the manuscript draft.
23
24

25
26 **Funding statement.** This trial is entirely supported by a Fondazione AIRC per la Ricerca sul
27 Cancro Investigator Grant ID 21761 to Gabriele Zoppoli.
28
29

30 **Competing interests.** The authors declare no conflict of interest for the present study.
31
32
33
34
35
36

37 **Figure legends**

38
39
40
41 **Figure 1. Study diagram.** Blood and urine samples will be collected from patients yielding a
42 radiological breast lesion ≤ 2 cm with no evidence of lymph node disease (radiological T1N0).
43 Images and samples acquired from patients with stage I-IIA (T1N0 or T2N0 or T1N1a
44 neoplasia) BC at surgery will be analysed for diagnostic purposes together with images and
45 samples acquired from patients yielding benign breast lesions and from patients with negative
46 mammography. Blood and urine samples will be re-collected from all patients yielding invasive
47 neoplasia at diagnosis after surgery at the first oncologic visit, and will be analysed for the
48 prediction of BC recurrence.
49
50
51
52
53
54
55
56
57
58
59
60

Figure 2. Sample size diagram. Approximately 1,500 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.

Biometrics	Voluptuous habits	Endocrine status	Predisposition	Comorbidities
Height	Does/Did the patient smoke?	Is the patient in the pre- or post menopause 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 sostitutive 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?		

Tab. 1. Items investigated at the recruitment. Along with items concerning the inclusion and exclusion criteria, all patients will be asked about information of interest for the assessment of BC risk and possible interference with the evaluation of circulating biomarkers.

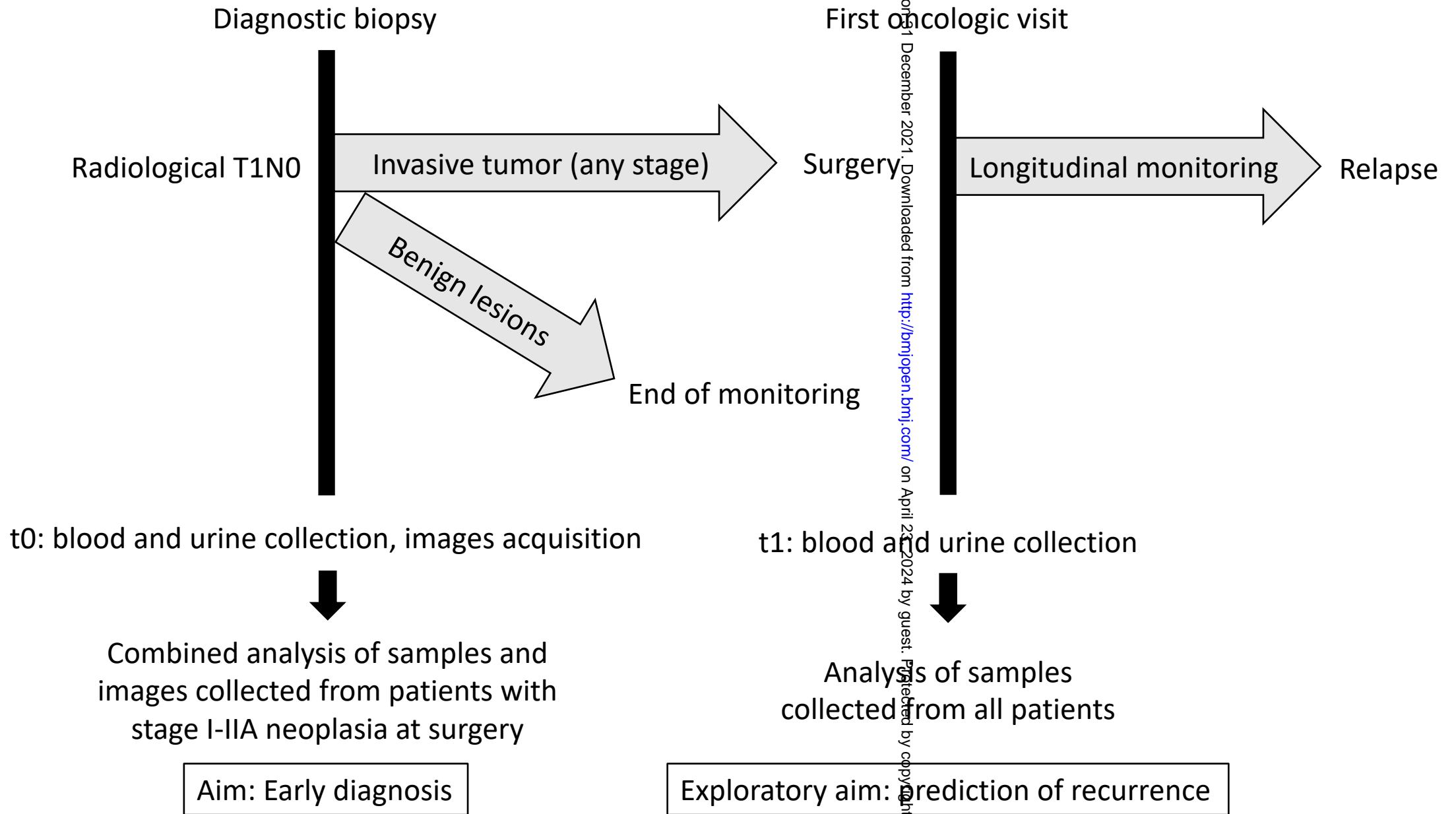
References

- 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. doi:10.1148/radiol.2017161519
- 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. doi:10.1200/JCO.2015.63.4147
- 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. doi:10.1038/ncomms15081
- 4 Bettgowda 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. doi:10.1126/scitranslmed.3007094
- 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. doi:10.1126/science.aar3247
- 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. doi:10.1016/j.ccell.2015.09.018
- 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. doi:10.1016/j.canlet.2018.02.002

- 1
2
3 8 Shen SY, Singhanian R, Fehringer G, *et al.* Sensitive tumour detection and classification
4 using plasma cell-free DNA methylomes. *Nature* 2018;**563**:579–83.
5
6 doi:10.1038/s41586-018-0703-0
7
8
9
10 9 Kalluri R. The biology and function of exosomes in cancer. *J Clin Invest* 2016;**126**:1208–
11 15. doi:10.1172/JCI81135
12
13
14
15 10 Valdora F, Houssami N, Rossi F, *et al.* Rapid review: radiomics and breast cancer.
16 *Breast Cancer Res Treat* 2018;**169**:217–29. doi:10.1007/s10549-018-4675-4
17
18
19
20 11 Bhawal R, Oberg AL, Zhang S, *et al.* Challenges and Opportunities in Clinical
21 Applications of Blood-Based Proteomics in Cancer. *Cancers* 2020;**12**.
22 doi:10.3390/cancers12092428
23
24
25
26
27 12 Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: A
28 comprehensive review. *Clin Genet* 2019;**95**:643–60. doi:10.1111/cge.13514
29
30
31
32 13 Aravanis AM, Lee M, Klausner RD. Next-Generation Sequencing of Circulating Tumor
33 DNA for Early Cancer Detection. *Cell* 2017;**168**:571–4. doi:10.1016/j.cell.2017.01.030
34
35
36
37 14 Bennett NC, Farah CS. Next-generation sequencing in clinical oncology: next steps
38 towards clinical validation. *Cancers* 2014;**6**:2296–312. doi:10.3390/cancers6042296
39
40
41
42 15 Stirzaker C, Taberlay PC, Statham AL, *et al.* Mining cancer methylomes: prospects and
43 challenges. *Trends Genet* 2014;**30**:75–84. doi:10.1016/j.tig.2013.11.004
44
45
46
47 16 Salomon MP, Orozco JIJ, Wilmott JS, *et al.* Brain metastasis DNA methylomes, a novel
48 resource for the identification of biological and clinical features. *Sci Data*
49 2018;**5**:180245. doi:10.1038/sdata.2018.245
50
51
52
53 17 Anfossi S, Babayan A, Pantel K, *et al.* Clinical utility of circulating non-coding RNAs - an
54 update. *Nat Rev Clin Oncol* 2018;**15**:541–63. doi:10.1038/s41571-018-0035-x
55
56
57
58
59
60

- 1
2
3 18 Koh W, Pan W, Gawad C, *et al.* Noninvasive in vivo monitoring of tissue-specific global
4 gene expression in humans. *Proc Natl Acad Sci U S A* 2014;**111**:7361–6.
5
6 doi:10.1073/pnas.1405528111
7
8
9
10
11 19 AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine
12 through an International Consortium. *Cancer Discov* 2017;**7**:818–31. doi:10.1158/2159-
13 8290.CD-17-0151
14
15
16
17 20 Nuzzo PV, Berchuck JE, Korthauer K, *et al.* Detection of renal cell carcinoma using
18 plasma and urine cell-free DNA methylomes. *Nat Med* 2020;**26**:1041–3.
19
20 doi:10.1038/s41591-020-0933-1
21
22
23
24 21 Nassiri F, Chakravarthy A, Feng S, *et al.* Detection and discrimination of intracranial
25 tumors using plasma cell-free DNA methylomes. *Nat Med* 2020;**26**:1044–7.
26
27 doi:10.1038/s41591-020-0932-2
28
29
30
31
32 22 Shen SY, Burgener JM, Bratman SV, *et al.* Preparation of cfMeDIP-seq libraries for
33 methylome profiling of plasma cell-free DNA. *Nat Protoc* 2019;**14**:2749–80.
34
35 doi:10.1038/s41596-019-0202-2
36
37
38
39 23 Webber J, Stone TC, Katilius E, *et al.* Proteomics analysis of cancer exosomes using a
40 novel modified aptamer-based array (SOMAscan™) platform. *Mol Cell Proteomics*
41 2014;**13**:1050–64. doi:10.1074/mcp.M113.032136
42
43
44
45
46 24 Raffield LM, Dang H, Pratte KA, *et al.* Comparison of Proteomic Assessment Methods in
47 Multiple Cohort Studies. *Proteomics* 2020;**20**:e1900278. doi:10.1002/pmic.201900278
48
49
50
51 25 Lambin P, van Stiphout RGPM, Starmans MHW, *et al.* Predicting outcomes in radiation
52 oncology--multifactorial decision support systems. *Nat Rev Clin Oncol* 2013;**10**:27–40.
53
54 doi:10.1038/nrclinonc.2012.196
55
56
57
58 26 Dietterich TG. Machine-learning research; four current directions. <https://go.gale.com> ›
59
60

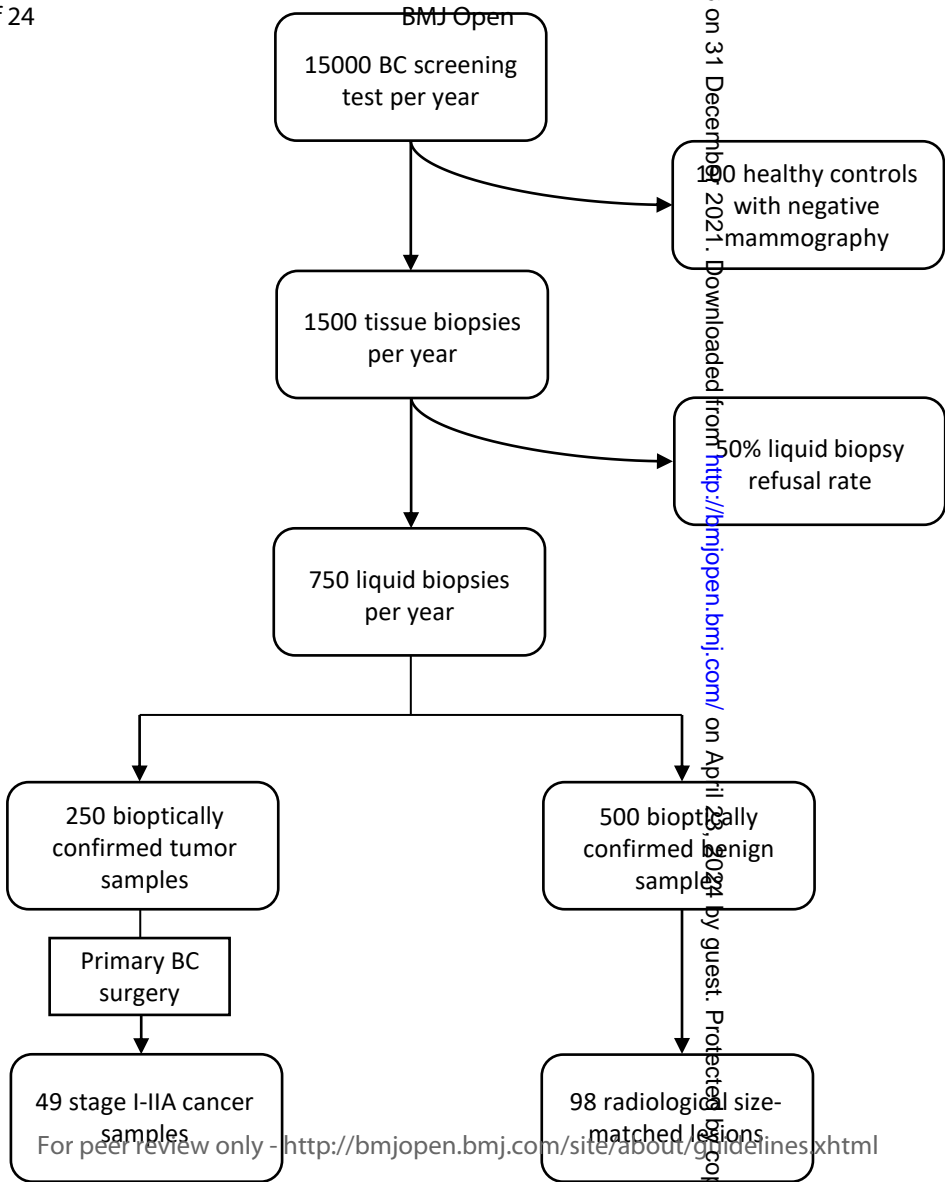
- 1
2
3 *i.do*<https://go.gale.com> > *i.do*
4
5 1997;**18**:97+.[https://go.gale.com/ps/i.do?id=GALE%7CA20475202&sid=googleScholar&](https://go.gale.com/ps/i.do?id=GALE%7CA20475202&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=07384602&p=AONE&sw=w)
6
7 [v=2.1&it=r&linkaccess=abs&issn=07384602&p=AONE&sw=w](https://go.gale.com/ps/i.do?id=GALE%7CA20475202&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=07384602&p=AONE&sw=w)
8
9
10
11 27 Thrun S, Pratt L. Learning to Learn: Introduction and Overview. In: Thrun S, Pratt L,
12 eds. *Learning to Learn*. Boston, MA: : Springer US 1998. 3–17. doi:10.1007/978-1-
13 4615-5529-2_1
14
15
16
17 28 Dobbin KK, Simon RM. Sample size planning for developing classifiers using high-
18 dimensional DNA microarray data. *Biostatistics* 2007;**8**:101–17.
19
20 doi:10.1093/biostatistics/kxj036
21
22
23
24 29 Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a
25 classifier for microarray data? *Clin Cancer Res* 2008;**14**:108–14. doi:10.1158/1078-
26 0432.CCR-07-0443
27
28
29
30
31 30 Qiu J, Xu J, Zhang K, *et al*. Refining Cancer Management Using Integrated Liquid
32 Biopsy. *Theranostics* 2020;**10**:2374–84. doi:10.7150/thno.40677
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



054256 on 11 December 2021. Downloaded from <http://bmjopen.bmj.com/> on April 20, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36



56 on 31 December 2021. Downloaded from <http://bmjopen.bmj.com/> on April 28, 2024 by guest. Protected by copyright.

BMJ Open

Development of a hoRizontal data intEgration classifier for NOinvasiVe early diAgnosis of breasT cancer: the RENOvATE study protocol

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-054256.R1
Article Type:	Protocol
Date Submitted by the Author:	29-Oct-2021
Complete List of Authors:	Ravera, Francesco; Università degli Studi di Genova, Department of internal Medicine Cirmena, Gabriella; Università degli Studi di Genova, Department of internal Medicine Dameri, Martina; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia Gallo, Maurizio; Università degli Studi di Genova, Department of internal Medicine Vellone, Valerio; Università degli Studi di Genova, Department of Surgical Sciences and Integrated Diagnostic Fregatti, Piero; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia Friedman, Daniele; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia Calabrese, Massimo; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia Ballestrero, Alberto; University of Genoa, Department of Internal Medicine Tagliafico, Alberto; Università degli Studi di Genova, Department of Health Sciences Ferrando, Lorenzo; Ospedale Policlinico San Martino Istituto di Ricovero e Cura a Carattere Scientifico per l'Oncologia Zoppoli, Gabriele; Università degli Studi di Genova, Department of internal Medicine
Primary Subject Heading:	Oncology
Secondary Subject Heading:	Diagnostics, Radiology and imaging
Keywords:	Breast tumours < ONCOLOGY, Breast imaging < RADIOLOGY & IMAGING, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Cancer genetics < GENETICS, Adult oncology < ONCOLOGY



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3 **Development of a hoRizontal data intEgration classifier for NOinvasiVe early**
4 **diAgnosis of breasT cancEr: the RENOVATE study protocol**
5
6
7
8

9 Authors: Francesco Ravera¹, Gabriella Cirmena¹, Martina Dameri², Maurizio Gallo²,
10 Valerio Gaetano Vellone^{2,4}, Piero Fregatti^{2,4}, Daniele Friedman^{2,4}, Massimo Calabrese²,
11 Alberto Ballestrero^{1,2} Alberto Tagliafico^{2,3,*}, Lorenzo Ferrando^{2,*}, Gabriele Zoppoli^{1,2*,#}
12
13
14
15
16
17
18
19

20 ¹Department of Internal Medicine and Medical Specialties DiMI, University of Genoa,
21 16132, Italy
22
23

24 ²IRCCS Ospedale Policlinico San Martino, 16132 Genoa, Italy
25

26 ³Department of Health Sciences DISSAL, University of Genoa, 16132, Italy
27

28 ⁴Department of Surgical Sciences and Integrated Diagnostic DISC, 16132 Genoa, Italy
29
30
31
32

33 *These Authors equally contributed to the present manuscript
34

35 # Correspondence should be addressed to Gabriele Zoppoli: gabriele.zoppoli@unige.it
36
37
38
39
40

41 **Keywords:** breast cancer, early diagnosis, liquid biopsy, circulating biomarkers, radiomics
42

43 **Word count:** 4640
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 tumor 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.

Registration: ClinicalTrials.gov, 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 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 (i.e. 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

1. 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 characterization 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 suspected region considering the genotypic and phenotypic heterogeneity of BCs [3], (c) long turnaround time for recall tests, even in high-level centers. In particular, the suboptimal sensitivity of screening procedures leads to non-diagnosed tumors 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].

1
2
3 The compelling necessity of increasing the accuracy of screening and diagnostic procedures
4 for cancer has brought to a relevant advancement in the pursuit of accurate biomarkers able
5 to efficiently detect and characterize it. To date, however, no protocol with effective and
6 recognized clinical validity for the early diagnosis of BC has been developed.
7
8

9
10
11 Amongst the most promising biomarkers, circulating tumor DNA (ctDNA) and cell-free
12 methylated DNA (cfMeDNA), exosomes, and microRNA (miRNA) have been the subject of
13 relevant scientific reports [4–9]. Moreover, machine-learning (so called “deep learning”)
14 algorithms have been applied to traditional radiology imaging techniques for diagnosis
15 assistance with exciting results in the field of radiomics [10]. Results in the proteomics field
16 have instead been few, in part due to the cumbersome methodologies often applied to the
17 study of these molecules [11].
18
19

20 Since it is hardly conceivable that a single biomarker is able to achieve 100% accuracy in the
21 early detection of BC, the primary aim of the present study is to merge for this purpose the
22 assessment of multiple biological analytes with the refinement of radiomics algorithms,
23 overcoming the aforementioned limitations in terms of accuracy of the individual biomarkers.
24 The concept of combining different data layers to reach a better classifier compared to the
25 individual analytes is referred to as horizontal data integration (HDI) classification. Therefore,
26 the overall goal of the project is to develop an HDI classifier enabling early noninvasive
27 diagnosis of BC with similar accuracy compared to breast biopsies.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 **2. Methods and analysis**

46 **2.1. Study design**

47 This is a prospective translational case-control study with the primary aim of assessing the
48 clinical validity of several biomarkers, individually and combined, in the early detection of BC
49 in a real-life clinical setting (see Figure 1 for study diagram). Patients are recruited at the
50 presentation to the Diagnostic Senology Unit of Ospedale Policlinico San Martino (Genoa,
51 Italy) for the execution of the breast biopsy after the radiological detection of a suspect breast
52 lesion ≤ 2 cm (i.e. radiological T1, BIRADS-3/4/5) with no radiological evidence of axillary or
53
54
55
56
57
58
59
60

1
2
3 distant disease. These patients are asked, upon completion of an informed written consent, to
4 donate four peripheral blood tubes (30 mL total) and one urine sample (40 mL). Of these, only
5 samples collected from patients with stage I-IIA neoplasia at surgery (T1N0 or T2N0 or T1N1a)
6 or histologically confirmed benign lesions will be analysed. A second sample is obtained from
7 those patients with confirmed histological diagnosis of invasive BC at the first oncologic visit
8 after breast primary surgery (t1) according to normal practice, independently from possible
9 neoadjuvant treatments. No limitations based on tumor stage will be made for the analysis of
10 the latter. Blood and urine samples are also collected and analysed from a cohort of 100
11 healthy women with two consecutive negative mammograms (BIRADS-1 or BIRADS-0 with
12 negative ecography) as well.

13
14 All patients are screened with the following inclusion and exclusion criteria. Inclusion criteria
15 include: written informed consent; breast lesions detected by digital bilateral mammography;
16 age ≥ 18 years and ≤ 75 years; eligibility for diagnostic biopsy (tru-cut or Vacuum Assisted
17 Breast Biopsy) as per normal clinical practice for study population, or absence of breast
18 lesions at the digital mammography for healthy controls (BIRADS-1 or BIRADS-0 with negative
19 ultrasound); ability and willfulness to comply with the protocol requirements.

20
21 Exclusion criteria include: previous history of invasive cancer, any type; clinical or radiological
22 suspicion of advanced or metastatic cancer at the time of screening; known history of active
23 or treated autoimmune or manifest chronic or seasonal and active allergic disorders (with the
24 exception of autoimmune thyroiditis); history of major trauma or surgery during the 24 weeks
25 before screening; history of active infectious disease, either chronic or acute occurring during
26 the 8 weeks before screening; history of known acute or chronic cardiac, kidney, or liver
27 disease disorders or acute cardiac events.

2.2. Candidate biomarkers

2.2.1. Circulating tumor DNA and cell-free Methylated DNA

28
29 ctDNA obtained via liquid biopsy has shown serious potential not only in the early diagnosis
30 of cancer but *inter alia* as an effective marker for its recurrence, longitudinal monitoring and
31 response to therapy [12]. Current methods for detecting ctDNA however are mostly based on
32

1
2
3 sequencing somatic mutations from cell-free DNA (cfDNA), a process constrained by relevant
4 limitations in clinical applicability due to (a) expectable low sensitivity in early stage cancers
5 given the limited number of recurrent mutations in ctDNA [8,13], (b) the vast heterogeneity of
6 DNA mutations occurring in a single tumor together with nonspecific mutations profiles along
7 different patients and cancer types [12,13], (c) current cost-prohibitive impact of cfDNA Next
8 Generation Sequencing (NGS) for mutation assessment [14]. On the other hand, the
9 assessment of cfMeDNA may overcome the limitations outlined above. DNA methylation is
10 commonly involved in cellular development, tissue-specific gene expression, and regulation
11 of imprinted alleles, with widespread effects on cellular growth and genomic stability [15].
12 Changes in methylomes in cancer are usually associated with alterations of the transcriptional
13 outcome and genomic instability, priming or enhancing carcinogenesis. Given the impact of
14 methylation changes on cellular equilibrium, different methylomes may be associated with
15 specific biological features providing useful information for the early diagnosis of cancer [16].
16 Shen *et al.* recently developed a sensitive, immunoprecipitation-based protocol to analyze the
17 methylomes of small quantities of cfDNA, providing an efficient method to detect large-scale
18 DNA methylation changes that are enriched for tumor-specific patterns [8]. cfMeDNA
19 immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) was able to find cancer-
20 derived DNA methylation events in cfDNA and to distinguish multiple cancer types from
21 healthy controls, besides effectively discriminating different methylation patterns across
22 diverse cancers.

23 24 25 **2.2.2. RNA-based biomarkers**

26 RNA-based biomarkers can be referred either to coding or non-coding RNAs. Non-coding
27 RNAs, including miRNAs and long non coding RNAs, have been extensively studied over the
28 last years as promising markers for cancer early diagnosis and monitoring, being of particular
29 interest for their stability and quantity in the bloodstream [17]. On the other hand, coding-RNAs
30 include the cell-free messenger RNA, whose evaluation may allow the non-invasive
31 assessment of the whole-body transcriptome [18] with relevant implications for several clinical
32 purposes.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

2.2.3. Circulating proteins

Proteomic analyses in neoplastic patients have been evaluated in regard to diverse clinical purposes with suboptimal results, given the challenging characterization of circulating proteins. The assessment of alterations in protein-mediated signaling 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].

2.3. Sample collection

2.3.1. Peripheral blood and urine collection, processing and storage

One investigator, a research nurse, and two biologists are collaborating in the first eighteen months of the project to collect at t0 and t1 two PAXgene® Blood ccfDNA Tubes CE-IVD (PreAnalytix, 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, PBMCs, and total RNA. PAXgene tubes are first centrifuged for 15 minutes at 1900 rcf at room temperature (RT), then the collected plasma is further centrifuged for 10 minutes at 1900 rcf at RT. EDTA tubes are centrifuged for 15 minutes at 1600 rcf at RT and the collected plasma is further centrifuged at 1900 rcf for 10 minutes 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 stabilize 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 minutes, 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, 3 Compartments.

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

2.3.2. Resources

A 150 sqm 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

1
2
3 extraction, quality control, amplification, library preparation and sequencing happen in
4 separate, clean environments optimized for such tasks. Every operation, from sample
5 processing, to nucleic acid extraction and sequencing, is semi-automated in order to minimize
6 cross-sample contamination and ensure optimal reproducibility and consistency of operations.
7
8 A dedicated freezer is available in the laboratories for sample storage. The laboratory is
9 equipped with a ThermoFisher Scientific Ion™ S5 XL sequencer. Data collection and the dry-
10 lab part of the analyses are performed on high performance workstations maintained in high-
11 security, dedicated environments to avoid the risk of data loss and sensible data cyber-theft.
12
13
14
15
16
17
18
19

20 **2.4. Analysis and integration of circulating biomarkers with radiomics**

21 **2.4.1. Ultrasensitive NGS on ctDNA**

22 We designed a tagged-amplicon NGS panel covering 101 short regions of frequently mutated
23 genes in BC. For this task, we analyzed the GENIE dataset v3.0 [19], which includes 3,269
24 sequenced primary and metastatic BC. This dataset was chosen for its size and homogeneity.
25 Since our purpose was looking at the most frequent, non-overlapping (i.e. mutually exclusive)
26 mutations, the sample size of GENIE was deemed as adequate for the panel design. Amongst
27 the most common cancer genes in BC, very few behave as dominant oncogenes (i.e. with
28 frequent mutation in limited regions of the gene, like PIK3CA), whereas several behave as
29 tumor suppressors (i.e., with rare mutations occurring along the whole gene, like CDH1). By
30 dividing all the genes sequenced by GENIE into 9,603 regions shorter than 50 base pairs, we
31 realized that to detect at least one mutation, we would need to sequence only 101 regions to
32 cover 80% of the whole GENIE dataset, but 420 regions to increase such detection rate to
33 88%. In particular, amplicons covering TP53 (30 exonic regions), CDH1 (50 exonic regions),
34 GATA3 (13 exonic regions), and PIK3CA (7 exonic regions) would optimize the coverage by
35 minimizing the size of our panel. TP53 is planned to be analyzed, together with CDH1, for all
36 its exons. On the other hand, PIK3CA presents mostly hotspot mutations in exons 9 and 20.
37 Hence, only those two exons are going to be covered by this design. The 101 selected regions
38 will be assessed in plasma ctDNA using a custom design, which leverages the novel,
39 proprietary tagged-amplicon OncoMine™ cfDNA methodology by ThermoFisher Scientific Inc.
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Its use allows for a limit of detection of 0.1% at 20,000x sequencing coverage with 20 ng of
4 circulating DNA, well attainable with two peripheral blood samples collected in PAXgene
5 tubes. ThermoFisher defines ctDNA positivity as the detection of three molecular families.
6
7 However, this threshold is arbitrary and will be adapted depending on the actual class the
8 samples fall in, in order to optimize the accuracy of the multiomic HDI. cfDNA extraction is
9 performed by using the QIAamp Mini Elute cfDNA Mini or Midi Kit (Qiagen, Hilden, Germany)
10 on the Qiacube system, according to the Manufacturer's instructions. cfDNA is extracted from
11 at least 4 mL of plasma collected in the EDTA tubes and 2 mL of plasma collected in the
12 PAXgene tubes, managing to obtain the requested amount of cfDNA for cfMeDIP-seq (~ 5 ng)
13 and the assessment of ctDNA mutations (~15 ng). cfDNA is processed on an Ion Chef™ fluidic
14 handler, and sequenced using the tagged-amplicon methodology with our custom-design
15 panel using as per manufacturer's specifications on a Ion S5™ sequencer. Tertiary analysis
16 will be performed on a dedicated Ion Server System available in our laboratory. The library
17 size of our custom panel is half of the commercial ThermoFisher Scientific Ion Torrent™
18 OncoMine™ Comprehensive cfDNA™ panel, from which the aforementioned performance
19 considerations have been derived. Hence, the use of our custom-design panel appears
20 feasible and should perform better in light of its smaller size and subsequent deeper attainable
21 sequencing coverage.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

41 **2.4.2. Methylome profiling of cfDNA**

42
43 Recent works pointed out encouraging results in terms of accuracy of cfMeDIP-seq in the
44 detection of informative methylation changes of small quantities (1-10 ng) of plasma cfDNA
45 for the diagnosis of renal cell carcinoma and intracranial tumors [20,21], with the former being
46 effectively detected also by cfMeDIP-seq on urinary cfDNA. According to the protocol
47 elaborated by Shen *et al.*, cfMeDIP involves four steps: 1) cfDNA end repair, A-tailing and
48 adapter ligation; 2) cfMeDIP immunoprecipitation and enrichment using an antibody targeting
49 5 methylcytosine; 3) library preparation; 4) high throughput NGS on an Illumina platform for
50 cfMeDNA data [22]. This approach based on immunoprecipitation allows to avoid cfDNA
51
52
53
54
55
56
57
58
59
60

1
2
3 bisulfite treatment, typically used to study DNA methylation but burdened by the high rate of
4 DNA degradation during the process.
5
6

7 **2.4.3. Proteomics analysis**

8
9 The proposed task is based on the relative quantification of blood proteins by a novel, highly
10 multiplexed proteomic assay (SomaScan™) [23,24]. SomaScan™ (SomaLogic™) is
11 considered the most comprehensive protein array available so far for the relative quantification
12 of proteins. Over 7,500 proteins can be simultaneously analyzed starting from 55 µL of plasma.
13
14 This technology involves a new type of aptamers, which are single stranded DNA molecules
15 able to bind proteins, called SOMAmers. In addition to their high affinity for individual proteins,
16 SOMAmers have a unique 40-nucleotide sequence tag and a fluorescent label that allows
17 their identification and quantification in high-density microarrays. SOMAmers have been
18 successfully assembled in a commercial product allowing the comparative evaluation of
19 proteins in as low as 55µl of serum or plasma (or other biological fluids). The SOMAscan™
20 assay is a highly multiplexed, sensitive, quantitative, and reproducible proteomic tool for
21 biomarker discovery and development. SomaScan™ can detect proteins within a range of 10
22 logarithms, allowing for an unmatched sensitivity to detect even femtomolar protein
23 concentrations. The analysis of the SomaScan™ is performed using classic DNA array data
24 analysis and is based on bioinformatics tools that have been developed for gene array
25 analysis.
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 **2.4.4. Radiomics analyses**

44
45 A preliminary radiomics classifier has been developed by the Diagnostics Senology team of
46 Ospedale Policlinico San Martino, based on digital breast tomosynthesis (DBT) images of
47 consecutive participants from the ASTOUND trial [2]. Radiomics analyses were performed on
48 all DBT images within manually selected regions of interest (ROIs) including all the dense
49 parts of the breast and excluding the fatty parts. ROIs were selected by a single radiologist,
50 with proven expertise in quantitative image analysis. Descriptors of the preliminary classifier
51 were selected after initial screening of 104 radiomics features to reduce the risk of over-fitting
52 and according to features previously used to associate breast parenchymal patterns with
53
54
55
56
57
58
59
60

1
2
3 cancer risk [25]. For the present project, image features will be extracted from the same cases
4 and matched controls for whom NGS and proteomics analyses are performed, using an open
5 source software platform for medical image informatics, and will make use of advanced deep
6 learning methodologies.
7
8
9

10 11 **2.4.5. HDI classifier**

12 Sensitivity and specificity of ctDNA analysis, cfMeDNA tests, proteomic analysis, and
13 radiomics for the early detection of BC will be assessed. Data from these classifiers will be
14 subsequently processed and used to generate the HDI model, based on the ensemble
15 learning approach methodology. Ensemble learning combines predictions of multiple
16 individual classifiers obtained by different techniques such as random forest, support vector
17 machine or general linear modeling in order to enhance generalization power [26], avoid
18 overfitting, and increase the strength and reliability of the final outcome [27]. Specifically, the
19 outcomes from ctDNA, cfMeDNA, proteomics and radiomics tests will be combined by using
20 a weighted-majority voting approach implemented in the R environment (caret package). We
21 do not envisage RNA-based classifiers to be initially included in our multianalyte model.
22
23
24
25
26
27
28
29
30
31
32
33
34

35 **2.4.6. Experimental validation and other analyses**

36 Upon completion of these experiments, we will seek to validate the possible discovery of novel
37 biomarkers and to facilitate their transfer to clinical applicability. Other promising noninvasive
38 biomarkers will be studied, thanks to the unique sample set at our disposal. We foresee the
39 possibility to perform exosome-enriched miRNA sequencing, PBMC transcriptome
40 sequencing by using the Ampliseq™ Transcriptome solution on our Ion S5™ XL sequencer.
41 Eventually, such analyses may be integrated into the HDI classifier to augment the accuracy
42 of BC early diagnosis.
43
44
45
46
47
48
49
50

51 **2.5. Study outcomes**

52 This study aims at assessing the performance of multiple analytes, individually and combined
53 in a HDI classifier, in the early detection of BC. The primary outcome is defined upon the
54 histopathological diagnosis of early invasive BC characterized as per SIAPEC/ASCO/CAP
55 criteria, with radiological extension ≤ 2 cm (radiological T1) and stage I-IIA at surgery (T1N0
56
57
58
59
60

1
2
3 or T2N0 or T1N1a). Benign lesions are defined upon the detection of radiological lesions ≤ 2
4 cm without the presence of invasive neoplasia at the first biopsy and at surgery, if performed.
5
6
7 *In situ* tumors are included in the latter group.
8

9 **2.6. Clinical data**

10
11 Electronic case report forms have been designed for the annotation of patients' clinical data.
12
13 These include study participants' demographics, biometric parameters such as height, weight
14 and BMI, assumption of alcohol and smoke, information about the endocrine status (pre- or
15 post-menopause, age at menarche, assumption of endocrine therapy, number of
16 pregnancies), familiarity for BC and predisposing mutations, comorbidities. Items included in
17 the eCRF are provided in **Tab. 1**. Tumor histology and immunohistochemical features will be
18 annotated as well.
19
20
21
22
23
24
25

26 **2.7. Data security and confidentiality**

27
28 Patients' pseudo-anonymized radiological images as well as demographic and
29 anatomopathological data are collected according to the local ethics committee guidelines
30 using a dedicated, state-of-the-art firewalled data collection system, OpenClinica®, hosted by
31 University of Genova servers.
32
33
34
35

36 **2.8. Statistical analysis**

37 **2.8.1. Projected sample size**

38
39 The Diagnostics Senology Unit of San Martino Hospital is the highest-level referral center in
40 Italy for a population basin of more than 2,000,000 people. In a single year, approximately
41 15,000 mammograms are performed (Figure 2). Of these, 1,500 yield a radiologic suspicion
42 of malignancy with an ensuing biopsy. Assuming a 50% refusal rate by our patients to undergo
43 trial-specific blood collection, we foresee to enroll 750 patients with radiologically suspect
44 small breast lesions. Among these patients, we assume diagnosis of invasive BC will be
45 confirmed in one third. Of these 250 bioptic diagnoses of BC, which will undergo primary
46 surgery, as per internal historical records, 40% are assumed to be pT1 and 60% higher than
47 pT1. Assuming another 10% failure rate in sample processing and storage, we will have the
48 potential to collect imaging data and samples from approximately 90 patients with pT1 BC and
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 180 radiological size-matched lesions in a single year (Figure 2). Patient enrollment is
4 projected in a time frame of 12 months from the beginning of the study, whereas samples
5 collection at t1 is expected to end approximately 6 months after the enrollment of the last
6 patient.
7
8
9

10 11 **2.8.2. Sample size calculation**

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

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

24 It may be possible that, adopting HDI classification, the actual needed sample size will be
25 smaller than calculated. At present however, we are not aware of well-established statistical
26 approaches to obtain a more robust estimation of the number needed for our experiments
27 other than the one we adopted.
28
29

30 **2.9. Amendments**

31 The original protocol for this study underwent three amendments, mainly aimed at including
32 the assessment of urinary biomarkers and expanding the cohort of analysed patients.
33
34

35 **2.9.1. Amendment #1: collection and analysis of urine samples**

36 The first amendment to this protocol, presented on February 18th 2020, involved the collection
37 and analysis of urine samples in addition to blood samples, which was not included in the
38 original version. An efficient method for the early diagnosis of BC based on the assessment
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 of urinary biomarkers would substantially eliminate the invasiveness of the overall procedure,
4 possibly facilitating the execution of large-scale screening campaigns.

7 **2.9.2. Amendment #2: healthy controls**

9 The second amendment, presented on February 4th 2021, involved the recruitment and
10 analysis of a cohort of 100 healthy women with negative mammography (BIRADS-1 or
11 BIRADS-0 with adjunct negative ultrasound) as a healthy control group, in addition to patients
12 undergoing breast biopsy. As the presence of benign breast lesions may determine possible
13 variations in blood and urinary analytes, especially in case of *in situ* tumors, the purpose of
14 the present amendment is to possibly categorize patients upon the presence or absence of
15 non-malignant breast lesions, enhancing the accuracy of the HDI classifier.
16
17

24 **2.9.3. Amendment #3: redefinition of the study cohort**

26 The third and final amendment, presented on May 27th 2021, involved several points.
27 According to the original protocol, concerning the cohort of patients with histological diagnosis
28 of invasive BC, only samples collected at t0 from patients with T1N0 tumor at surgery would
29 have been analysed. Moreover, the collection and analysis of t1 samples would have involved
30 only this group of patients, with the exclusive aim of distinguishing between tumor- and host-
31 specific molecular alterations in connection with the presence/absence of BC. With the present
32 amendment, samples collected at t0 will be analysed from patients assessed as with stage I-
33 IIA tumor at surgery (T1N0 or T2N0 or T1N1a), allowing an effective expansion of the sample
34 size, while remaining in the setting of early stage BC. Furthermore, the collection and analysis
35 of t1 samples will include all patients diagnosed as with invasive BC and sampled at t1, without
36 limitations in terms of tumor stage assessed at surgery. The analysis of t1 samples, coupled
37 with patients' longitudinal monitoring performed as per normal clinical practice, will allow the
38 assessment of the accuracy of a multi-analyte evaluation for the prediction of BC recurrence,
39 added as exploratory aim of the present study.
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

56 **2.10. Study current status**

57 The recruitment phase of the research project started on January 18th 2021. To date (October
58 21st 2021), 183 patients undergoing diagnostic biopsy were recruited. Of these, 98 presented
59
60

1
2
3 a benign lesion, while 51 presented a malignant lesion. Thirty-four patients were ruled out. Of
4 the recruited patients with the histological diagnosis of BC, 35 underwent breast surgery. Of
5 these, 23 patients had a stage I-IIA breast tumor. One hundred and ten patients with negative
6 mammography were recruited as healthy controls. Of these, 20 patients were ruled out. The
7 recruitment phase of the study will end by March 2022. Proteomics, ctDNA, and cfMeDIP-seq
8 experiments will be performed during the second year of the research project. The third year
9 of the research project will be dedicated to assess the performance of the individual classifiers,
10 including the one based on radiomics. During the fourth year we will proceed to the wet lab
11 validation, besides performing the transcriptomic experiments. The fifth and final year of the
12 research project will be dedicated to build and optimize the HDI classifier.

2.11. Patient and Public Involvement

24
25
26 There was no patient or public involvement in the design of this study.

3. Ethics and dissemination

27
28
29
30
31
32
33 Written informed consents are obtained from each study participant. Participant information
34 sheet includes the main information of the study protocol, the known side effects and any risk
35 implied in the participation to the study, beside the contact information of study investigators.
36
37 All data are deidentified and no patient-related information will be revealed during analysis.
38
39 The Regione Liguria Ethics Committee c/o Ospedale Policlinico San Martino has approved
40 the study (reference number: 2019/75, study ID: 4452).

41
42
43 All information concerning patients included in this study are covered by strict confidentiality
44 in compliance with the General Data Protection Regulation EU 2016/679 (GDPR) and D.lgs.
45 30.06.2003, n. 196, as modified from D.lgs. 10.08.2018, n. 101. The study is conducted in
46 accordance with the national law and according to international guidelines for the conduction
47 of clinical trials referred to as "Good Clinical Practice".

48
49
50 Results will be published in international peer-reviewed scientific journals.

4. DISCUSSION

1
2
3 The assessment of circulating biomarkers for clinical purposes is currently burdened by
4 several issues, including suboptimal accuracy and the lack of standardized pre-analytical and
5 analytical procedures. Overcoming these limitations would allow the achievement of minimally
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

The assessment of circulating biomarkers for clinical purposes is currently burdened by several issues, including suboptimal accuracy and the lack of standardized pre-analytical and analytical procedures. Overcoming these limitations would allow the achievement of minimally invasive and personalized 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 enrollment, patients undergo a rigorous selection in order to avoid possible confounding factors that can affect the assessment of tumor-specific 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 tumor 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

1
2
3 this protocol, given the typical transition of *in situ* BC to invasive cancer and the consequent
4 necessity of its early detection and eradication.
5

6
7 However, the protocol outlined above presents relevant advantages compared to other studies
8 aimed at the same purpose, including the application of recent and cutting-edge techniques
9 to a selected but realistic cohort of patients, possibly bringing to effective advancements in
10 current standards BC patients management, besides directing future researches in the
11 panorama of translational medicine applied to oncology.
12
13
14
15
16

17
18
19
20 **Author statement.** GZ conceived the study. FR wrote the manuscript. GC and MD provided
21 a significant contribution in the section "Methods and analysis" . MG, VGV, PF, DF, MC, AB,
22 AT, and LF critically reviewed the manuscript draft.
23
24

25
26 **Funding statement.** This trial is entirely supported by a Fondazione AIRC per la Ricerca sul
27 Cancro, Investigator Grant ID 21761 to Gabriele Zoppoli.
28
29

30 **Competing interests.** The authors declare no conflict of interest for the present study.
31
32
33
34
35
36

37 **Figure legends**

38
39 **Figure 1. Study diagram.** Blood and urine samples will be collected from patients yielding a
40 radiological breast lesion ≤ 2 cm with no evidence of lymph node dissemination (radiological
41 T1N0). Images and samples acquired from patients with stage I-IIA (T1N0 or T2N0 or T1N1a
42 neoplasia) BC at surgery will be analysed for diagnostic purposes together with images and
43 samples acquired from patients yielding benign breast lesions and from patients with negative
44 mammography. Blood and urine samples will be re-collected from all patients yielding invasive
45 neoplasia at diagnosis after surgery at the first oncologic visit, and will be analysed for the
46 prediction of BC recurrence.
47
48
49
50
51
52
53
54

55
56
57
58 **Figure 2. Sample size diagram.** Approximately 1,500 breast biopsies per year are performed
59 at the Diagnostics Senology Unit of San Martino Hospital. Of a projected number of 750 liquid
60

1
2
3 biopsies, we foresee to collect samples and acquire mammograms from at least 49 patients
4 with stage I-IIA BC, and 98 patients with radiological size-matched lesions, along with those
5 samples and images acquired from 100 healthy women with two consecutive negative
6 mammograms.
7
8
9
10
11
12
13

Biometrics	Voluptuous habits	Endocrine status	Predisposition	Comorbidities
Height	Does/Did the patient smoke?	Is the patient in the pre- or post menopause 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?		

14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Tab. 1. Items investigated at the recruitment. Along with items concerning the inclusion and exclusion criteria, information of interest for the assessment of BC risk and possible interference with the evaluation of circulating biomarkers will be annotated.

References

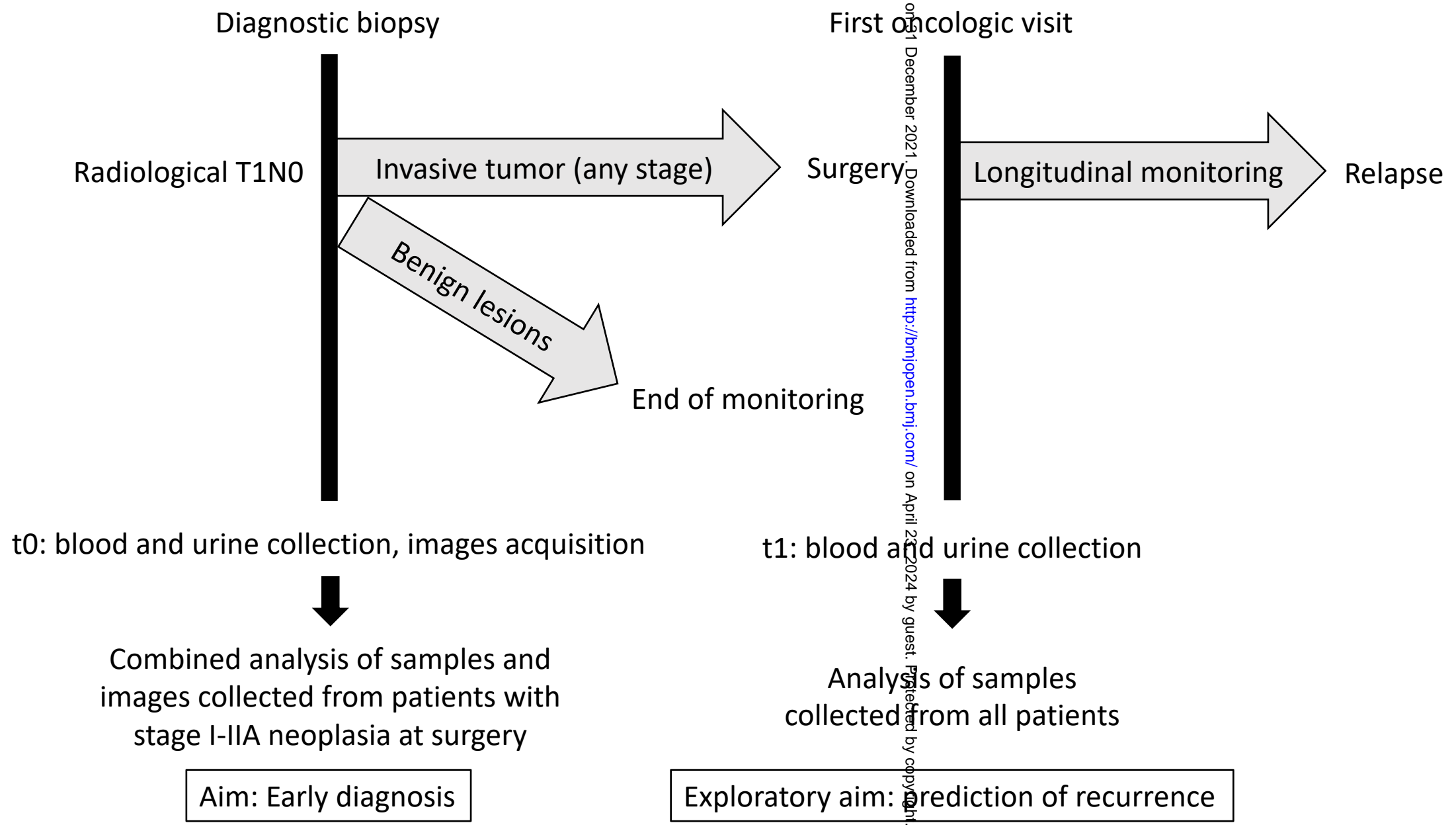
- 1 Sprague BL, Arao RF, Miglioretti DL, *et al.* National Performance Benchmarks for
2 Modern Diagnostic Digital Mammography: Update from the Breast Cancer Surveillance
3 Consortium. *Radiology* 2017;**283**:59–69. doi:10.1148/radiol.2017161519
4
5
6
7
8
9
- 10 2 Tagliafico AS, Calabrese M, Mariscotti G, *et al.* Adjunct Screening With Tomosynthesis
11 or Ultrasound in Women With Mammography-Negative Dense Breasts: Interim Report
12 of a Prospective Comparative Trial. *J Clin Oncol* 2016;**34**:1882–8.
13
14
15
16
17 doi:10.1200/JCO.2015.63.4147
18
19
- 20 3 Chung W, Eum HH, Lee H-O, *et al.* Single-cell RNA-seq enables comprehensive tumour
21 and immune cell profiling in primary breast cancer. *Nat Commun* 2017;**8**:15081.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
- 56 4 Bettgowda C, Sausen M, Leary RJ, *et al.* Detection of circulating tumor DNA in early-
57 and late-stage human malignancies. *Sci Transl Med* 2014;**6**:224ra24.
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
- 1 Shen SY, Singhanian R, Fehring G, *et al.* Sensitive tumour detection and classification
2 using plasma cell-free DNA methylomes. *Nature* 2018;**563**:579–83.
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

- 1
2
3 9 Kalluri R. The biology and function of exosomes in cancer. *J Clin Invest* 2016;**126**:1208–
4 15. doi:10.1172/JCI81135
5
6
7
8 10 Valdora F, Houssami N, Rossi F, *et al*. Rapid review: radiomics and breast cancer.
9 *Breast Cancer Res Treat* 2018;**169**:217–29. doi:10.1007/s10549-018-4675-4
10
11
12
13 11 Bhawal R, Oberg AL, Zhang S, *et al*. Challenges and Opportunities in Clinical
14 Applications of Blood-Based Proteomics in Cancer. *Cancers* 2020;**12**.
15 doi:10.3390/cancers12092428
16
17
18
19
20 12 Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: A
21 comprehensive review. *Clin Genet* 2019;**95**:643–60. doi:10.1111/cge.13514
22
23
24
25 13 Aravanis AM, Lee M, Klausner RD. Next-Generation Sequencing of Circulating Tumor
26 DNA for Early Cancer Detection. *Cell* 2017;**168**:571–4. doi:10.1016/j.cell.2017.01.030
27
28
29
30 14 Bennett NC, Farah CS. Next-generation sequencing in clinical oncology: next steps
31 towards clinical validation. *Cancers* 2014;**6**:2296–312. doi:10.3390/cancers6042296
32
33
34
35 15 Stirzaker C, Taberlay PC, Statham AL, *et al*. Mining cancer methylomes: prospects and
36 challenges. *Trends Genet* 2014;**30**:75–84. doi:10.1016/j.tig.2013.11.004
37
38
39
40 16 Salomon MP, Orozco JIJ, Wilmott JS, *et al*. Brain metastasis DNA methylomes, a novel
41 resource for the identification of biological and clinical features. *Sci Data*
42 2018;**5**:180245. doi:10.1038/sdata.2018.245
43
44
45
46
47
48 17 Anfossi S, Babayan A, Pantel K, *et al*. Clinical utility of circulating non-coding RNAs - an
49 update. *Nat Rev Clin Oncol* 2018;**15**:541–63. doi:10.1038/s41571-018-0035-x
50
51
52
53 18 Koh W, Pan W, Gawad C, *et al*. Noninvasive in vivo monitoring of tissue-specific global
54 gene expression in humans. *Proc Natl Acad Sci U S A* 2014;**111**:7361–6.
55 doi:10.1073/pnas.1405528111
56
57
58
59
60

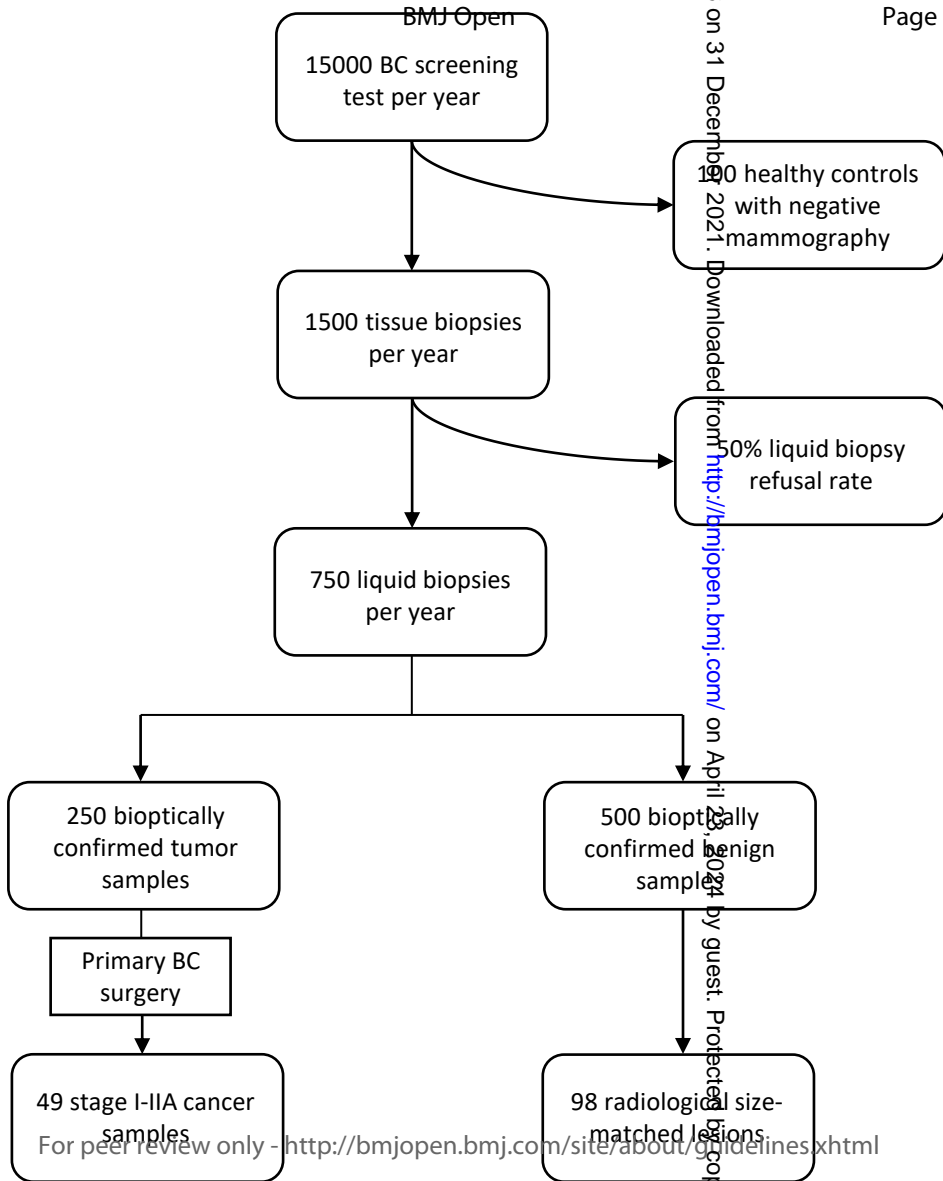
- 1
2
3 19 AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine
4 through an International Consortium. *Cancer Discov* 2017;**7**:818–31. doi:10.1158/2159-
5 8290.CD-17-0151
6
7
8
9
10 20 Nuzzo PV, Berchuck JE, Korthauer K, *et al*. Detection of renal cell carcinoma using
11 plasma and urine cell-free DNA methylomes. *Nat Med* 2020;**26**:1041–3.
12 doi:10.1038/s41591-020-0933-1
13
14
15
16
17 21 Nassiri F, Chakravarthy A, Feng S, *et al*. Detection and discrimination of intracranial
18 tumors using plasma cell-free DNA methylomes. *Nat Med* 2020;**26**:1044–7.
19 doi:10.1038/s41591-020-0932-2
20
21
22
23
24 22 Shen SY, Burgener JM, Bratman SV, *et al*. Preparation of cfMeDIP-seq libraries for
25 methylome profiling of plasma cell-free DNA. *Nat Protoc* 2019;**14**:2749–80.
26 doi:10.1038/s41596-019-0202-2
27
28
29
30
31 23 Webber J, Stone TC, Katilius E, *et al*. Proteomics analysis of cancer exosomes using a
32 novel modified aptamer-based array (SOMAscan™) platform. *Mol Cell Proteomics*
33 2014;**13**:1050–64. doi:10.1074/mcp.M113.032136
34
35
36
37
38
39 24 Raffield LM, Dang H, Pratte KA, *et al*. Comparison of Proteomic Assessment Methods in
40 Multiple Cohort Studies. *Proteomics* 2020;**20**:e1900278. doi:10.1002/pmic.201900278
41
42
43
44 25 Lambin P, van Stiphout RGPM, Starmans MHW, *et al*. Predicting outcomes in radiation
45 oncology--multifactorial decision support systems. *Nat Rev Clin Oncol* 2013;**10**:27–40.
46 doi:10.1038/nrclinonc.2012.196
47
48
49
50
51 26 Dietterich TG. Machine-learning research; four current directions. <https://go.gale.com> ›
52 [i.dohttps://go.gale.com](https://go.gale.com) › [i.do](https://go.gale.com)
53
54 1997;**18**:97+. [https://go.gale.com/ps/i.do?id=GALE%7CA20475202&sid=googleScholar&](https://go.gale.com/ps/i.do?id=GALE%7CA20475202&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=07384602&p=AONE&sw=w)
55
56
57
58
59
60

- 1
2
3 27 Thrun S, Pratt L. Learning to Learn: Introduction and Overview. In: Thrun S, Pratt L,
4 eds. *Learning to Learn*. Boston, MA: : Springer US 1998. 3–17. doi:10.1007/978-1-
5 4615-5529-2_1
6
7
8
9
10
11 28 Dobbin KK, Simon RM. Sample size planning for developing classifiers using high-
12 dimensional DNA microarray data. *Biostatistics* 2007;**8**:101–17.
13 doi:10.1093/biostatistics/kxj036
14
15
16
17 29 Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a
18 classifier for microarray data? *Clin Cancer Res* 2008;**14**:108–14. doi:10.1158/1078-
19 0432.CCR-07-0443
20
21
22
23
24 30 Qiu J, Xu J, Zhang K, *et al*. Refining Cancer Management Using Integrated Liquid
25 Biopsy. *Theranostics* 2020;**10**:2374–84. doi:10.7150/thno.40677
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

054256 on 11 December 2021. Downloaded from <http://bmjopen.bmj.com/> on April 20, 2024 by guest. Protected by copyright.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

56 on 31 December 2021. Downloaded from <http://bmjopen.bmj.com/> on April 28, 2024 by guest. Protected by copyright.