BMJ Open

Molecular markers of risk of subsequent invasive breast cancer in women with ductal carcinoma in situ: protocol for a population-based cohort study

Thomas E Rohan ,[1] Mindy Ginsberg ,[1] Yihong Wang ,[2,3] Fergus J Couch ,[4,5] Heather S Feigelson ,[6] Robert T Greenlee ,[7] Stacey Honda ,[8] Azadeh Stark ,[9,10] Dhananjay Chitale ,[9,10] Tao Wang ,[1] Xiaonan Xue ,[1] Maja H Oktay ,[11,12] Joseph A Sparano ,[13] Olivier Loudig [14]

To cite: Rohan TE, Ginsberg M, Wang Y, *et al*. Molecular markers of risk of subsequent invasive breast cancer in women with ductal carcinoma in situ: protocol for a population-based cohort study. *BMJ Open* 2021;**11**:e053397. doi:10.1136/bmjopen-2021-053397

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-053397).

Received 11 May 2021
Accepted 06 October 2021

For numbered affiliations see end of article.

Correspondence to
Dr Thomas E Rohan;
thomas.rohan@einsteinmed.org

ABSTRACT

Introduction Ductal carcinoma in situ (DCIS) of the breast is a non-obligate precursor of invasive breast cancer (IBC). Many DCIS patients are either undertreated or overtreated. The overarching goal of the study described here is to facilitate detection of patients with DCIS at risk of IBC development. Here, we propose to use risk factor data and formalin-fixed paraffin-embedded (FFPE) DCIS tissue from a large, ethnically diverse, population-based cohort of 8175 women with a first diagnosis of DCIS and followed for subsequent IBC to: identify/validate miRNA expression changes in DCIS tissue associated with risk of subsequent IBC; evaluate ipsilateral IBC risk in association with two previously identified marker sets (triple immunopositivity for p16, COX-2, Ki67; Oncotype DX Breast DCIS score); examine the association of risk factor data with IBC risk.

Methods and analysis We are conducting a series of case–control studies nested within the cohort. Cases are women with DCIS who developed subsequent IBC; controls (2/case) are matched to cases on calendar year of and age at DCIS diagnosis. We project 485 cases/970 controls in the aim focused on risk factors. We estimate obtaining FFPE tissue for 320 cases/640 controls for the aim focused on miRNAs; of these, 173 cases/346 controls will be included in the aim focused on p16, COX-2 and Ki67 immunopositivity, and of the latter, 156 case–control pairs will be included in the aim focused on the Oncotype DX Breast DCIS score®. Multivariate conditional logistic regression will be used for statistical analyses.

Ethics and dissemination Ethics approval was obtained from the Institutional Review Boards of Albert Einstein College of Medicine (IRB 2014-3611), Kaiser Permanente Colorado, Kaiser Permanente Hawaii, Henry Ford Health System, Mayo Clinic, Marshfield Clinic Research Institute and Hackensack Meridian Health, and from Lifespan Research Protection Office. The study results will be presented at meetings and published in peer-reviewed journals.

Strengths and limitations of this study

- This is a large, multicentre population-based cohort study of women with ductal carcinoma in situ (DCIS) of the breast designed to identify the associations of clinical, epidemiological and pathological factors, and of molecular changes in DCIS tissue, with risk of subsequent invasive breast cancer.
- The study will have considerable statistical power.
- Although some individuals may be misclassified with respect to exposure (ie, their status with respect to the results of the molecular assays), the study methods are highly sensitive and reproducible, and laboratory staff are blinded to case–control status.
- Generalisation of the study findings to other populations will require caution, but cases and controls are selected independently of molecular markers, so it is unlikely that unavailability or unsuitability of tissue for analysis will bias the study results.

INTRODUCTION

Ductal carcinoma in situ (DCIS), which arises in the terminal duct lobular unit of the breast,[1] is considered to be a non-obligate precursor of invasive breast cancer (IBC).[2] About 5%–14% of patients diagnosed with DCIS and treated with breast-conserving therapy, with or without radiation, experience an ipsilateral IBC and 1%–6% experience a contralateral IBC over a period of 10 years.[3–5] Overall, compared with women in the general population, women with a history of DCIS have anywhere from a 1.5-fold to a 10-fold increase in risk of subsequent IBC.[3 6–9]

The introduction of screening mammography has led to a substantial increase in the detection of DCIS over the past 2–3 decades.[2] However, the treatment of DCIS remains variable, and many DCIS patients are either undertreated or overtreated.[10 11] In this regard, there is a critical need for elucidation of the molecular differences between lesions that are associated with risk of IBC

development and those that are not, as this may help not only to reduce the occurrence of IBC, but also to prevent overtreatment of patients with relatively low risk of IBC.[12 13]

Gene expression profiling and immunohistochemical analysis of IBCs have led to improvements in prediction of their clinical behaviour.[14–16] Given that DCIS is associated with increased risk of IBC development,[10] we hypothesise that study of molecular changes in DCIS may similarly help to identify novel biomarkers that predict IBC risk. In this regard, there is some evidence that the Oncotype DX Breast DCIS score®, derived from a multigene expression assay that incorporates genes related to proliferation, as well as *PR* and *GSTM1*, and five reference genes, predicts risk of subsequent ipsilateral IBC.[10 17] Similarly, triple positivity for immunohistochemically detected expression of p16, COX-2 and Ki67 has also been associated with increased risk of ipsilateral IBC.[18] However, these findings require confirmation. Furthermore, we hypothesise that novel prognostic (and ultimately predictive) markers may emerge from assessment in DCIS tissue of gene expression patterns on a global scale. Of relevance here are microRNAs (miRNAs), small regulatory noncoding RNAs that control gene expression[19 20] and that, when dysregulated, contribute to the development of breast cancer.[21 22]

Against this background, our overarching goal is to facilitate the detection of patients with DCIS at risk of IBC development. To this end, we propose to use clinical, epidemiological and histopathological data, and archival formalin-fixed paraffin-embedded (FFPE) tissue, from a large, ethnically diverse, population-based multicentre cohort of 8175 patients initially diagnosed with DCIS in community-based hospitals and health plans and followed for subsequent IBC development. Our specific aims are to identify/validate miRNA expression changes in DCIS tissue associated with risk of subsequent IBC, to evaluate ipsilateral IBC risk in association with two previously identified marker sets (triple immunopositivity for p16, COX-2, Ki67; the Oncotype DX Breast DCIS score), and to examine the association of clinical, epidemiological and histopathological variables with IBC risk. Furthermore, we will store genomic DNA (extracted simultaneously with RNA) for subsequent study of somatic genetic changes related to risk of IBC.

METHODS

Study population and design

We assembled a multicentre population-based cohort of 8175 women aged ≥18 years, with no history of IBC, who received a first histological diagnosis of DCIS between 12 January 1987 and 20 December 2016. The study started on 15 September 2017 and will continue until 31 July 2023 at the earliest. The cohort was constructed by including DCIS patients from six large community-based integrated healthcare delivery systems across the USA: Henry Ford Health System (Detroit, Michigan, USA), Kaiser Permanente (KP) Colorado (Denver, Colorado, USA), KP Hawaii (Honolulu, Hawaii, USA), Marshfield Clinic (Marshfield, Wisconsin, USA), Mayo Clinic (Rochester, Minnesota, USA) and Montefiore Medical Center (Bronx, New York, USA) (table 1). Collectively, the participating centres provide comprehensive medical care for patients who are broadly representative of the racial/ethnic and socioeconomic distribution of the underlying populations.

Table 1 Cohort numbers, overall and by centre

Site	Calendar period of DCIS diagnosis	No of subjects with DCIS	No of subsequent cases of invasive breast cancer*
Henry Ford	1991–2015	1737	149
KP Colorado	2006–2016	1154	33
KP Hawaii	1987–2016	936	65
Marshfield	1990–2016	1233	66
Mayo	1988–2016	1358	108
Montefiore	1994–2016	1757	64
Total	1987–2016	8175	485

*For follow-up until 2020.
DCIS, Ductal carcinoma in situ.

Cohort follow-up and ascertainment of IBC cases

Using the extensive electronic medical record (EMR) and administrative databases at each institution, patients are followed passively from the date of their DCIS diagnosis until the date of development of subsequent IBC, death, health plan disenrollment or the end of follow-up, whichever comes first. In each centre, the occurrence of subsequent IBC (ipsilateral or contralateral) in the DCIS cohort is ascertained by linking records from the cohort to the respective tumour registry and/or to other EMR data. Given that all participating centres provide integrated care, all healthcare procedures for patients are typically conducted within the health systems, and any outside services are captured through reimbursement/claims data. This integrated model also facilitates the collection of tumour tissue, which usually is stored centrally. The number of IBCs ascertained in each centre and the calendar period of the DCIS diagnoses are shown in table 1.

Nested case–control studies

We are conducting our investigation as a series of case–control studies nested in the DCIS cohort.

Cases are women with a first diagnosis of DCIS and with a subsequent diagnosis of IBC at least 6 months after the index DCIS diagnosis (table 1) (women treated by mastectomy who develop subsequent chest wall IBC are not included in the study). Controls are women with a first diagnosis of DCIS, with no history of breast cancer, with no history of bilateral mastectomy prior to the date of diagnosis of IBC for the corresponding case,

and who were alive but had not developed IBC during the same follow-up period as that for the corresponding case. For each of the DCIS subjects who developed a subsequent IBC (cases), we select two controls using risk-set sampling. Each control is individually matched to the corresponding case on calendar year of (mostly within ±1 year) and age at (generally ±1 year) diagnosis of DCIS. We have limited the number of matching variables so as to retain flexibility in the analysis. Nevertheless, we are measuring potentially relevant variables (eg, treatment, adjuvant endocrine therapy) and will be able to adjust for such variables in the analysis. We expect to include 485 cases and 970 controls in the aim focused on clinical/epidemiological/histopathological variables and IBC risk. We project that we will obtain FFPE tissue for 320 of the cases and 640 controls, and these subjects will be included in the aim focused on miRNAs; of these, 173 cases/346 controls will be included in the aim focused on p16, COX-2 and Ki67 immunopositivity, and of the latter, 156 case–control pairs will be included in the aim focused on the Oncotype DX Breast DCIS score.

Risk factor data

We use the EMR to extract data on variables such as age at and year of DCIS diagnosis; family history of breast cancer; height and weight; method of DCIS detection (palpation, mammography); tumour laterality; tumour size; treatment (including breast conserving surgery, radiotherapy and hormonal therapy); menstrual and reproductive history; use of oral contraceptives and hormone replacement therapy; and cigarette and alcohol consumption. The data are being obtained in a standardised manner at all participating institutions by using a chart abstraction manual to guide data abstraction; chart abstractors are trained centrally in use of the manual. Risk factor data are being abstracted for all cases and their matched controls, regardless of whether DCIS tissue is obtained.

Tissue blocks/sections

For the cases and matched controls included in this study, we attempt to retrieve FFPE DCIS tissue blocks from the pathology archives of the participating institutions. To date, DCIS tissue blocks have been obtained for 72% of the cases; we attempt to obtain DCIS tissue for two controls per case, replacing potential controls for whom tissue is unobtainable. For cases with subsequent contralateral IBC (and their matched controls), four 5 µm sections (for oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2/neu (HER2/neu) and H&E), plus 16 ten µm sections (for RNA/DNA extraction), are cut from each FFPE block; for cases with subsequent ipsilateral IBC, **7** five µm sections are cut (for ER, PR, HER2, p16, COX-2, Ki67, H&E), in addition to 16 ten µm sections. The last of the 5 µm unstained sections (ie, the section after the sections taken for immunohistochemistry) undergoes H&E staining.

Experimental methods

All histopathology reviews and laboratory assays are performed blinded to the case-control status of the study subjects.

Histopathology/receptor status

H&E sections are reviewed and classified by the study pathologist (YW) according to standard criteria with respect to nuclear grade, architectural pattern, necrosis and microcalcification[23]; areas of DCIS are circled on the coverslips (women with DCIS with microinvasion are not included). The sections stained for ER, PR and HER2/neu are interpreted in accordance with American Society of Clinical Oncology/College of American Pathologists (ASCO-CAP) guidelines.[24 25] ER/PR positivity is defined as ≥1% cells staining positive[24] and HER2 positivity is defined as a score of 3+. For those subjects for whom tissue blocks cannot be located, we will use information on histology and receptor status abstracted from the medical records.

RNA/DNA extraction

Using the H&E section as a guide, the DCIS lesions are macrodissected from the 16 ten µm unstained sections and then subjected to co-extraction of RNA and DNA using our simultaneous TRIzol RNA/DNA extraction protocol.[26] Total RNA is quantified and visually evaluated on an Agilent Bioanalyzer (RNA Nanochip), and DNA is quantified on a Qubit instrument (dsDNA High sensitivity assay kit). Stock RNA samples are aliquoted for sequencing and qPCR experiments. RNA/DNA samples are stored at −80°C.

MiRNA expression profiling using next-generation sequencing

Using 100 ng of total RNA per sample, batches of 18 samples will be set up in individual ligation reactions (using T4 RNA ligase 2 truncated K227Q), using 18 different 3' adenylated barcoded adapters, at 4°C for 16 hours. We will include cases and their matched controls in the same libraries. Ten unique RNA calibrators (each 22 nucleotides long) without any match to human or mouse genomes will be spiked at 0.026 nM into each individual sample ligation, and will be used as quality and normalisation controls. As a further control, duplicate DCIS case and control samples selected from the different libraries will be analysed in three additional libraries. After 3' ligation of the barcoded adapters, the 18 reactions will be heat-deactivated, and the ligated products will be combined and size-selected on a 15% polyacrylamide gel (PAGE). The ligated products will be excised, purified and subjected to a global 5'adapter ligation. After ligation, the products will be size-selected on a 12% PAGE, purified and reverse-transcribed. The cDNA templates will be subjected to a pilot PCR for identification of the appropriate number of amplification cycles. A final large-scale PCR will be set up, size selected (2% agarose gel), and sequenced on an Illumina HiSeq 2500 sequencer. Individual libraries will be analysed by

aligning the reads to the human genome and separating the samples using the 18 different 3' barcoded adapters.[27]

Real-time polymerase chain reaction (qPCR) miRNA quantification for validation of miRNA sequencing findings

100 ng of total RNA will be used to validate expression of the top 14 differentially expressed miRNAs identified by sequencing. MiRNA quantification will be performed in two steps, a single multiplex RT step (14 miRNAs and 2 endogenous controls—RNU6B, RNU44) using Taqman® reverse transcription kits and Taqman® primer RT pool (16 genes; Applied Biosystems (ABI)), followed by individual qPCR reactions (16 genes analysed in triplicate) using Taqman® universal PCR mastermix (No AmpErase; ABI) and individual PCR primers for each of the 16 genes. Triplicate miRNA expression measures will be quantified on a QuantStudio 6-flex real-time qPCR instrument. The controls will be used for normalisation of miRNA quantification.[28] The case-control fold-change (FC) differences will be calculated using the $2^{(\Delta\Delta Ct)}$ formula.[28]

p16, COX-2 and Ki67 immunohistochemistry

Slides (5 µm sections) from batches of eight case–control triplets (cases with subsequent ipsilateral IBC) are deparaffinised, rinsed in graded alcohol, heated to 96°C for 20 min in 10 mM sodium citrate buffer (pH 6.0), and placed into a Dako Autostainer Plus for processing. Endogenous peroxidase activity is quenched using 3% hydrogen peroxide in 1xPBS for 10 min. Blocking is performed by incubating sections in 5% normal donkey serum with 2% BSA for 30 min. The primary antibodies for p16 antigen (Fisher Scientific antibody Clone D25, Catalog# MA5-17142) used at 1:5000, for Ki67 antigen (Dako antibody, Clone MIB-1,Catalog# M7240) used at 1:200, and for COX-2 antigen (Dako Clone antibody CX-294, Catalog# M361701-2) used at 1:200, all diluted in antibody diluent, are incubated for 1 hour. The sections are stained using a labelled polymer-HRP anti-mouse antibody (Dako Envision System-HRP (DAB)) for 30 min to localise the antibody bound to the antigen, with diaminobenzidine as the final chromogen. After washes, the sections are lightly counterstained with haematoxylin and mounted in Permount. All immunohistochemistry is performed separately for each antibody and thus requires three 5 µm sections/specimen. The p16, COX-2, and Ki67 stained sections are evaluated as described elsewhere.[18]

Oncotype DX DCIS score evaluation using Nanostring

Given cost considerations, it is not feasible to use the Oncotype DX Breast DCIS score® in this study. Therefore, as an alternative, expression of the 12 genes from the Oncotype DX DCIS assay plus three additional controls (MRPL19, COX-2, p16) and one internal control (18s rRNA) will be measured in parallel with expression of 770 other genes using the Nanostring nCounter PanCancer progression panel. To this end, 100 ng of total RNA from matched cases (with subsequent ipsilateral IBC) and controls will be analysed simultaneously. The data will be processed using the nSolver Software (Nanostring) and then subjected to further statistical analyses.

Statistical analysis

We will examine the associations between clinical, epidemiological and histopathological variables and risk of subsequent IBC using multivariate conditional logistic regression.

To identify/validate miRNA expression changes in DCIS tissue associated with risk of subsequent IBC, we are conducting a two-stage study. Based on our feasibility work, we estimate that we will obtain FFPE DCIS tissue for at least 320 cases and 640 controls. Of these subjects, 200 cases/400 controls will be included in the discovery stage and 120 cases/240 controls in the validation stage. For analysis of the sequencing data (discovery stage), after quality control and normalisation,[29 30] we will first explore the expression data using heat maps with unsupervised hierarchical clustering via Spearman distance and complete linkage as well as multidimensional scaling plots. This will be followed by supervised analyses to compare miRNA levels in the cases and controls using regression analysis, for example, limma in Bioconductor,[31 32] which allows for adjustment for covariates and paired comparisons. As a complementary approach, we will model the read counts as a negative binomial distribution to account for discrete sequence data.[33] We will rank differentially expressed miRNAs based on both statistical significance and degree of FC. These analyses will allow us to establish a list of candidate miRNAs for validation, while adjusting for clinical, epidemiological and histopathological factors.

To address missing information on risk factors, we will perform multiple imputation after examining the missingness mechanism.[34–37] The imputed risk factor data will be used together with the miRNAs to develop prediction models using the 'stack' method for combining multiply imputed data into one dataset with proper weights.[38] Sensitivity analyses will be performed to evaluate the robustness of models to imputation approaches. We will build our model by starting with clinical/epidemiological/histopathological factors and then perform joint analyses by adding candidate miRNAs to identify the most important miRNAs. For this purpose, we will use conditional logistic regression with elastic net penalities.[39 40]

Although it is impossible to accurately predict the number of miRNAs that will be associated with IBC development, we will choose the top 14 miRNAs for validation to ensure that we will have sufficient power for validation given our sample size.[41 42] We will perform similar analyses to those in the discovery stage to validate the association of each candidate miRNA with risk of IBC. Additionally, we will evaluate the potential clinical utility of the identified miRNAs from the final elastic regression model obtained in the discovery stage by calculating the area under the receiver-operating characteristic curve (AUC) and its SD.[43 44] The difference in AUCs between models with and without the identified miRNAs will be calculated

and the 95% CI will be estimated using a bootstrap procedure.[45] In additional analyses, we will explore the associations of miRNAs with IBC risk by side of IBC (ipsi- and contralateral) and by type of DCIS (using the St. Gallen criteria for molecular subtypes).[46]

We will evaluate the relationship between two previously examined marker sets (p16, COX-2, and Ki67 protein expression; Oncotype DX Breast DCIS score®) and ipsilateral IBC risk using conditional logistic regression. For p16, COX-2 and Ki67 protein expression, IBC risk will be estimated for subjects positive for all three markers (compared with all other subgroups combined).[18] In addition, we will explore risk for up to eight subgroups characterised by the three protein expression markers. For the Oncotype DX Breast DCIS DCIS score®, we will instead use the Nanostring nCounter PanCancer progression panel to measure expression of the relevant genes. Quality control and normalisation will be performed using nSolver Software (Nanostring). For the analysis of these data, we will first use similar statistical approaches to those reported in an earlier study.[10] Because the original DCIS score was calculated by giving fixed weights, the score may not be optimal. Therefore, we will also derive an alternative DCIS score by building a conditional logistic regression model that has the best predictive value from the data. To alleviate the problem of overfitting, 10-fold cross-validation will be used.[47] ROC analyses will be used to evaluate the improvement in risk prediction as a result of the addition of the biomarker score to the clinical, epidemiological and histopathological risk factors. AUCs with 95% CIs will be estimated and compared using the bootstrap approach.[48]

Statistical power

For the proposed analyses of clinical/epidemiological/histopathological variables, we project a sample size of 485 cases/970 controls. This will yield 80% power at α=0.05 to detect ORs between 1.17 and 1.20 at missing data rates of 0%–20% for a continuous risk factor, and ORs of 1.38–1.43 at missing rates between 0%–20% for categorical risk factors (assuming an exposure prevalence of 0.5).

For our work on miRNAs, we use the following formula to approximate the minimum FC required to achieve power (1-β) at statistical significance level (α): $[\text{Log}(\text{FC})]^2=(z_{1-\alpha/2}+z_{\beta})^2\ [(1/\mu+\sigma^2)/n_1(1/\mu+\sigma^2)/n_2]$, in which μ is read count, σ is the coefficient of variation of counts within a group, and n_1 and n_2 are sample sizes for cases and controls, respectively.[49] We considered a wide range of expression variation [low (σ=0.1) and high (σ=0.4)] and expression levels [low (μ=10), moderate (μ=50) and high (μ=100)]. Table 2 summarises the minimum FC required to achieve a given power (80%) for fixed sample sizes for the discovery and validation analyses. Even under the most conservative scenario considered [low expression level (read counts as low as 10); high variation within a group (0.4)], we will have 80% power to identify miRNAs with FCs>1.23 at an overly conservative Bonferroni significance level of 0.05/1000 in the discovery stage. For the validation stage, we will have similar power to validate markers with FCs>1.23 at a significance level of 0.0036, corresponding to a Bonferroni corrected p value of 0.05 for 14 independent markers. The FCs needed to achieve satisfactory power are realistic: our preliminary data showed that significant miRNAs had FCs in the range 1.23–5.8. The corresponding FCs for ipsilateral (54% of all cases) and contralateral cases are 1.34 and 1.37, respectively, magnitudes consistent with findings in our pilot work.

Table 2 The minimum fold-changes required to achieve 80% power for the miRNA analyses

Variation within a group (σ)	Expression level (μ)	Stage 1 (200 cases/400 controls; α=0.05/1000) minimum fold-change	Stage 2 (120 cases/240 controls; α=0.05/14) minimum fold-change
0.1	10	1.15	1.15
	50	1.08	1.08
	100	1.06	1.06
0.4	10	1.23	1.23
	50	1.20	1.19
	100	1.19	1.19

For the aim focused on p16, COX-2 and Ki67 protein expression, we expect that 173 of the 320 cases with DCIS tissue will be available, as we observe that ~54% of subsequent IBCs are ipsilateral. Assuming that the proportion of triple positive controls is ~9%,[10] and with two controls per case, we will have 80% power to detect an OR of 2.2 for the triple positive group versus the other groups combined, the same effect size as that observed in the previous study.[18]

For the aim focused on the Oncotype DX Breast DCIS score®, we expect 156 ipsilateral cases, given that in our pilot work, 90% of RNA extractions yielded sufficient RNA. Given a conservative assumption of 156 case–control pairs, we will have 80% power when α=0.05 to detect an OR of 2.01 per 50 point increase in DCIS score. This projected OR is smaller than the effect size observed previously.[10]

Patient and public involvement

No patient or public involvement.

ETHICS AND DISSEMINATION

Ethics approval was obtained from the Institutional Review Boards of the Albert Einstein College of Medicine (IRB 2014-3611), Kaiser Permanente Colorado, Kaiser Permanente Hawaii, the Henry Ford Health System, the Mayo Clinic, the Marshfield Clinic Research Institute, and Hackensack Meridian Health, and from the Research Protection Office of Lifespan. The results of this study will

be presented at national and international meetings and published in peer-reviewed journals.

DISCUSSION

Although women with DCIS are at increased risk of developing subsequent IBC compared with those without DCIS,[3 6–9] most women with DCIS do not go on to develop IBC. Furthermore, current clinical criteria do not discriminate well between those who will and will not develop IBC. Therefore, because the treatment of DCIS is generally aggressive, many women are treated unnecessarily.[11] This highlights the need for better markers of risk of IBC development. In this regard, we hypothesise that the molecular analysis of archival DCIS tissue will help to identify novel molecular markers associated with IBC risk, and may foster the development of risk stratification models and targeted therapies.[50]

To date, very few studies have addressed the identification of biomarkers associated with risk of subsequent IBC development in women with DCIS, largely because to do so requires long-term follow-up and almost inevitably entails use of archival specimens. Our molecular epidemiological study, which will apply state-of-the-art technologies to archival DCIS FFPE tissue for the detection of molecular changes associated with IBC development in a large, multicentre population-based cohort of women initially diagnosed with DCIS, has the potential to provide insight into the biology of DCIS and to lead to approaches that will help to differentiate between women who need and those who do not need enhanced surveillance and early aggressive treatment.

To investigate the associations of the Oncotype DX Breast DCIS score® and triple positivity for immunohistochemically detected expression of p16, COX-2 and Ki67 with risk of subsequent IBC, we will restrict attention to ipsilateral IBC, as per previous investigations in this area.[10 18] In contrast, for our work on miRNAs, we will include both ipsilateral and contralateral IBCs. We consider this to be appropriate because our study is motivated by evidence that breast cancer develops through the accumulation of molecular abnormalities (eg, structural chromosomal abnormalities, gene expression alterations) beginning in normal breast tissue, resulting from exposure to endogenous and environmental factors.[51 52] These exposures, beginning early in and continuing throughout life, can lead to molecular (and histological) changes in the tissue of both breasts, rendering them susceptible to the development of cancer,[51–53] as evidenced by: (1) the detection of cancer-predisposing molecular alterations in histologically normal breast tissue[51 54]; (2) shared molecular features in synchronous primary tumours, and concordance of hormone receptor expression status in synchronous and metachronous bilateral breast cancers[52 55–57]; (3) the observation that DCIS is often bilateral and multicentric[58] and (4) the increased risk of subsequent ipsilateral and contralateral IBC in women with DCIS.[3 6–9] Nevertheless, we will explore the associations of miRNAs with IBC risk by side of the subsequent IBC (ie, ipsilateral or contralateral).

Our study has a number of strengths. It was preceded by extensive pilot work demonstrating the feasibility of both obtaining clinical information and tissue blocks, and of performing the molecular assays using FFPE DCIS tissue. The study is large and will have considerable statistical power, and there is potential for further follow-up of the cohort to identify additional incident IBC cases.[59] The miRNA sequencing protocol that will be employed was developed for use with archival tissue up to 35 years old and allows for complete analysis of available transcripts in FFPE tissue[60 61]; we will validate the sequencing findings using Taqman qPCR assays, the gold-standard method. To address possible assay variability, the laboratory technicians undergo intensive training on the molecular assays. Furthermore, we use strict quality control for RNA/DNA extraction, preparation, and quantification, and we have built in experimental controls to monitor the accurate performance of our assays.[60 61] Importantly, the pathology and laboratory staff are blinded to case–control status.

The study also has some potential limitations. Although some individuals may be misclassified with respect to exposure (ie, their status with respect to the results of the molecular assays), the assays that we will use are highly sensitive and reproducible.[26 60–62] Also, although the study includes subjects from a wide range of populations across the USA, caution will be required in generalising the study findings. However, given that cases and controls will be selected independently of their status with respect to the molecular markers of interest, it is unlikely that unavailability or unsuitability of tissue for analysis will bias our results, and we expect the included subjects to be broadly representative of the underlying populations. Finally, a recent report showed that the 21-gene Oncotype DX breast recurrence score (RS) (Exact Sciences) was associated with increased risk of ipsilateral IBC (based on 19 incident, invasive cases).[63] Due to budgetary constraints, we are unable to incorporate the RS assay into our work. However, using the Nanostring assay, we will measure expression of the 21 genes in the RS assay[64] and will be able to examine the association between expression of these genes and IBC risk.

In conclusion, this study has the potential to provide insight into the pathogenesis of IBC by identifying molecular changes in DCIS lesions associated with altered IBC risk. Furthermore, the project has translational potential given that identification of molecular changes associated with increased IBC risk might enhance the clinical management of women with DCIS. Specifically, the study findings might lead both to the development of novel molecular screening modalities to identify women at increased risk of IBC and to approaches to prevention (eg, through risk stratification and tailored surveillance programmes, and through the development of novel targeted therapies).

Author affiliations
[1]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, USA
[2]Department of Pathology and Laboratory Medicine, Rhode Island Hospital and Lifespan Medical Center, Providence, Rhode Island, USA
[3]Warren Alpert Medical School of Brown University, Providence, Rhode Island, USA
[4]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA
[5]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[6]Institute for Health Research, Kaiser Permanente, Aurora, Colorado, USA
[7]Center for Clinical Epidemiology and Population Health, Marshfield Clinic Research Institute, Marshfield, Wisconsin, USA
[8]Center for Integrated Healthcare, Kaiser Permanente, Hawaii Permanente Medical Group, Honolulu, Hawaii, USA
[9]Department of Pathology and Laboratory Medicine, Henry Ford Health System, Detroit, Michigan, USA
[10]Breast Oncology Program and Department of Pathology, Henry Ford Health System, Detroit, Michigan, USA
[11]Department of Pathology, Albert Einstein College of Medicine/Montefiore Medical Center, Bronx, New York, USA
[12]Department of Anatomy and Structural Biology, Albert Einstein College of Medicine/Montefiore Medical Center, Bronx, New York, USA
[13]Department of Oncology, Albert Einstein College of Medicine/Montefiore Medical Center, Bronx, New York, USA
[14]Center for Discovery and Innovation, Hackensack Meridian Health, Nutley, New Jersey, USA

Acknowledgements The authors are grateful to the study subjects. Additionally, the authors thank the following individuals: Malka Felder, for her outstanding support in the development of the bioinformatics system for this project, and for data quality oversight; Minerva Manickchand for her dedication and for monitoring all workflow processes for the project; and DeeAnn Polacek, Terry Foss, Carla Rottscheit, Lisa Ott, and Syed Sajjad for their important contributions.

Contributors TER and OL conceived the study and secured funding for it. TER, OL, MG, YW, FJC, HSF, RTG, SH, AS, DC, TW, XX, MHO and JAS contributed to the refinement and finalisation of the study protocol. TER produced the initial draft of the manuscript and TER, OL, MG, YW, FJC, HSF, RTG, SH, AS, DC, TW, XX, MHO and JAS were involved in critical revision of the manuscript. and TER, OL, MG, YW, FJC, HSF, RTG, SH, AS, DC, TW, XX, MHO and JAS approved the submitted version of the manuscript.

Funding This study is supported by a grant from the National Cancer Institute/National Institutes of Health (R01CA218429). TER is supported by the Breast Cancer Research Foundation (BCRF-20–140).

Competing interests None declared.

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

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Thomas E Rohan http://orcid.org/0000-0001-6897-4045
Mindy Ginsberg http://orcid.org/0000-0002-4916-1273
Yihong Wang http://orcid.org/0000-0003-1252-5579
Fergus J Couch http://orcid.org/0000-0001-9417-9985
Heather S Feigelson http://orcid.org/0000-0001-6691-3740
Robert T Greenlee http://orcid.org/0000-0002-0618-7895
Stacey Honda http://orcid.org/0000-0001-5496-3759
Azadeh Stark http://orcid.org/0000-0002-4252-8263
Dhananjay Chitale http://orcid.org/0000-0002-4421-6431
Tao Wang http://orcid.org/0000-0003-0581-1251
Xiaonan Xue http://orcid.org/0000-0003-2872-8119
Maja H Oktay http://orcid.org/0000-0002-8115-3898
Joseph A Sparano http://orcid.org/0000-0002-9031-2010
Olivier Loudig http://orcid.org/0000-0001-5677-4716

REFERENCES

1 Allred DC, Mohsin SK, Fuqua SA. Histological and biological evolution of human premalignant breast disease. *Endocr Relat Cancer* 2001;8:47–61.
2 Coleman WB. Breast ductal carcinoma in situ: precursor to invasive breast cancer. *Am J Pathol* 2019;189:942–5.
3 Innos K, Horn-Ross PL. Risk of second primary breast cancers among women with ductal carcinoma in situ of the breast. *Breast Cancer Res Treat* 2008;111:531–40.
4 Li CI, Malone KE, Saltzman BS, *et al*. Risk of invasive breast carcinoma among women diagnosed with ductal carcinoma in situ and lobular carcinoma in situ, 1988-2001. *Cancer* 2006;106:2104–12.
5 Falk RS, Hofvind S, Skaane P, *et al*. Second events following ductal carcinoma in situ of the breast: a register-based cohort study. *Breast Cancer Res Treat* 2011;129:929–38.
6 Franceschi S, Levi F, La Vecchia C, *et al*. Second cancers following in situ carcinoma of the breast. *Int J Cancer* 1998;77:392–5.
7 Institute of Health and Welfare. *National breast and ovarian cancer centre. Risk of invasive breast cancer in women diagnosed with ductal carcinoma in situ in Australia between 1995 and 2005*. Canberra: Australian Institute of Health and Welfare, 2010.
8 Rawal R, Lorenzo Bermejo J, Hemminki K. Risk of subsequent invasive breast carcinoma after in situ breast carcinoma in a population covered by national mammographic screening. *Br J Cancer* 2005;92:162–6.
9 Wärnberg F, Yuen J, Holmberg L. Risk of subsequent invasive breast cancer after breast carcinoma in situ. *Lancet* 2000;355:724–5.
10 Solin LJ, Gray R, Baehner FL, *et al*. A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. *J Natl Cancer Inst* 2013;105:701–10.
11 Morrow M, Katz SJ. Addressing overtreatment in DCIS: what should physicians do now? *J Natl Cancer Inst* 2015;107:djv290.
12 Dowsett M, Goldhirsch A, Hayes DF, *et al*. International web-based consultation on priorities for translational breast cancer research. *Breast Cancer Res* 2007;9:R81.
13 Recht A, Rakovitch E, Solin LJ. Treatment and long-term risks for patients with a diagnosis of ductal carcinoma in situ. *JAMA Oncol* 2016;2:396.
14 Bombonati A, Sgroi DC. The molecular pathology of breast cancer progression. *J Pathol* 2011;223:307–17.
15 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
16 Carey LA, Perou CM, Livasy CA, *et al*. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 2006;295:2492–502.
17 Rakovitch E, Nofech-Mozes S, Hanna W, *et al*. Multigene expression assay and benefit of radiotherapy after breast conservation in ductal carcinoma in situ. *J Natl Cancer Inst* 2017;109. doi:10.1093/jnci/djw256. [Epub ahead of print: 01 Apr 2017].
18 Kerlikowske K, Molinaro AM, Gauthier ML, *et al*. Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis. *J Natl Cancer Inst* 2010;102:627–37.
19 Tavazoie SF, Alarcón C, Oskarsson T, *et al*. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* 2008;451:147–52.
20 Zhang X, Chen J, Radcliffe T, *et al*. An array-based analysis of microRNA expression comparing matched frozen and formalin-fixed paraffin-embedded human tissue samples. *J Mol Diagn* 2008;10:513–9.
21 Acunzo M, Romano G, Wernicke D, *et al*. MicroRNA and cancer-a brief overview. *Adv Biol Regul* 2015;57:1–9.
22 van Schooneveld E, Wildiers H, Vergote I, *et al*. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. *Breast Cancer Res* 2015;17:21.
23 Lester SC, Bose S, Chen Y-Y, *et al*. Protocol for the examination of specimens from patients with ductal carcinoma in situ of the breast. *Arch Pathol Lab Med* 2009;133:15–25.
24 Allison KH, Hammond MEH, Dowsett M, *et al*. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *J Clin Oncol* 2020;38:1346–66.
25 Wolff AC, Hammond MEH, Allison KH, *et al*. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *J Clin Oncol* 2018;36:2105–22.

26 Kotorashvili A, Ramnauth A, Liu C, *et al*. Effective DNA/RNA co-extraction for analysis of microRNAs, mRNAs, and genomic DNA from formalin-fixed paraffin-embedded specimens. *PLoS One* 2012;7:e34683.
27 Hafner M, Renwick N, Farazi TA, *et al*. Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods* 2012;58:164–70.
28 Giricz O, Reynolds PA, Ramnauth A, *et al*. Hsa-miR-375 is differentially expressed during breast lobular neoplasia and promotes loss of mammary acinar polarity. *J Pathol* 2012;226:108–19.
29 Hackenberg M, Sturm M, Langenberger D, *et al*. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 2009;37:W68–76.
30 Tarazona S, Furió-Tarí P, Turrà D, *et al*. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 2015;43:gkv711.
31 Kendziorski C, Irizarry RA, Chen K-S, *et al*. On the utility of pooling biological samples in microarray experiments. *Proc Natl Acad Sci USA* 2005;102:4252–7.
32 Kooperberg C, Aragaki A, Strand AD, *et al*. Significance testing for small microarray experiments. *Stat Med* 2005;24:2281–98.
33 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40.
34 Rubin D. *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons, 1987.
35 Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC Press, 1997.
36 Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci* 2007;8:206–13.
37 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377–99.
38 Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med* 2008;27:3227–46.
39 Sun H, Wang S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat Med* 2013;32:2127–39.
40 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Statistical Soc B* 2005;67:301–20.
41 Xiao Y, Hsiao T-H, Suresh U, *et al*. A novel significance score for gene selection and ranking. *Bioinformatics* 2014;30:801–7.
42 Yanofsky CM, Bickel DR. Validation of differential gene expression algorithms: application comparing fold-change estimation to hypothesis testing. *BMC Bioinformatics* 2010;11:63.
43 Cai T, Zheng Y. Non-parametric evaluation of biomarker accuracy under nested case-control studies. *J Am Stat Assoc* 2011;106:569–80.
44 Zhou QM, Zheng Y, Cai T. Assessment of biomarkers for risk prediction with nested case-control studies. *Clin Trials* 2013;10:677–9.
45 Efron B. *The jacknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics, 1987.
46 Zhou W, Jirström K, Amini R-M, *et al*. Molecular subtypes in ductal carcinoma in situ of the breast and their relation to prognosis: a population-based cohort study. *BMC Cancer* 2013;13:512.
47 Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res* 2008;14:5977–83.
48 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
49 Hart SN, Therneau TM, Zhang Y, *et al*. Calculating sample size estimates for RNA sequencing data. *J Comput Biol* 2013;20:970–8.
50 Sullenger BA, Nair S. From the RNA world to the clinic. *Science* 2016;352:1417–20.
51 Danforth DN, Jr.. Genomic changes in normal breast tissue in women at normal risk or at high risk for breast cancer. *Breast Cancer* 2016;10:109–46.
52 Lochhead P, Chan AT, Nishihara R, *et al*. Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. *Mod Pathol* 2015;28:14–29.
53 Botti C, Pescatore B, Mottolese M, *et al*. Incidence of chromosomes 1 and 17 aneusomy in breast cancer and adjacent tissue: an interphase cytogenetic study. *J Am Coll Surg* 2000;190:530–9.
54 Heaphy CM, Griffith JK, Bisoffi M. Mammary field cancerization: molecular evidence and clinical importance. *Breast Cancer Res Treat* 2009;118:229–39.
55 Coradini D, Oriana S, Mariani L, *et al*. Is steroid receptor profile in contralateral breast cancer a marker of independence of the corresponding primary tumour? *Eur J Cancer* 1998;34:825–30.
56 Newman LA, Sahin AA, Cunningham JE, *et al*. A case-control study of unilateral and bilateral breast carcinoma patients. *Cancer* 2001;91:1845–53.
57 Swain SM, Wilson JW, Mamounas EP, *et al*. Estrogen receptor status of primary breast cancer is predictive of estrogen receptor status of contralateral breast cancer. *J Natl Cancer Inst* 2004;96:516–23.
58 Alpers CE, Wellings SR. The prevalence of carcinoma in situ in normal and cancer-associated breasts. *Hum Pathol* 1985;16:796–807.
59 Elena JW, Travis LB, Simonds NI, *et al*. Leveraging epidemiology and clinical studies of cancer outcomes: recommendations and opportunities for translational research. *J Natl Cancer Inst* 2013;105:85–94.
60 Loudig O, Wang T, Ye K, *et al*. Evaluation and adaptation of a laboratory-based cDNA library preparation protocol for retrospective sequencing of archived microRNAs from up to 35-year-old clinical FFPE specimens. *Int J Mol Sci* 2017;18:E627.
61 Loudig O, Liu C, Rohan T, *et al*. Retrospective microRNA sequencing: complementary DNA library preparation protocol using formalin-fixed paraffin-embedded RNA specimens. *J Vis Exp* 2018. doi:10.3791/57471. [Epub ahead of print: 05 May 2018].
62 Hafner M, Renwick N, Brown M, *et al*. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* 2011;17:1697–712.
63 Rakovitch E, Sutradhar R, Nofech-Mozes S, *et al*. 21-gene assay and breast cancer mortality in ductal carcinoma in situ. *J Natl Cancer Inst* 2021;113:572–9.
64 Paik S, Shak S, Tang G, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.