Observational study of Age, test THreshold and frequency on English NAtional Mammography screening outcomes (ATHENA-M)

Protocol Version 1.4 16th February 2022

NCT05247463

Observational study of Age, test THreshold and frequency on English NAtional Mammography screening outcomes (ATHENA-M)

Protocol

Version 1.4 16th February 2022

Observational study of Age, test THreshold and frequency on English NAtional Mammography screening outcomes (ATHENA-M)

Protocol

Version 1.4 16th February 2022

Study Investigators

Sian Taylor-Phillips (Chief Investigator, University of Warwick)

Ros Given-Wilson (St Georges University Hospitals)

Matthew Wallis (Cambridge University Hospitals)

Louise Wilkinson (Oxford University Hospitals)

Sarah Pinder (Kings College London)

Jon Deeks (Birmingham University)

Alice Sitch (Birmingham University)

Jonathan Sterne (Bristol University)

Aileen Clarke (University of Warwick)

James Mason (University of Warwick)

David Jenkinson (University of Warwick)

Karoline Freeman (University of Warwick)

Olive Kearins (Public Health England)

Jackie Walton (Public Health England)

Sponsor: University of Warwick,

Identification Numbers

Office for Data Release: ODR1920_283 (pending)

Sponsor's Office Ref: SOC.03/20-21 (approved)

HRA: 21/LO/0120 (approved)

Breast Research Advisory Committee: BSP RAC 089 (Conditional approval)

ClinicalTrials.gov Identifier: NCT05247463

Protocol Previous versions

Version 1.0 January 2020 pre funding award.

Version 1.1 Changes from version 1.0 were as a result of feedback from NIHR funding panel, and include removal of work package 3 (analysis of mammographic abnormality type) and change of title. This version was sent to HRA for initial approvals.

Version 1.2 Minor amendments requested by Public Health England Office for Data Release. Page 7-8, expanded description of how the database differs from the POSTBOx study, and exact definition of the included cohort added. Data tables moved from appendix 2 to a separate file, appendix 2 adjusted to a broader description of the variables. 'Mortality and births Information System' corrected to 'civil registrations mortality file'. Primary source of breast cancer mortality outcome changes from cancer registry to civil registrations mortality file. David Jenkinson, Olive Kearins, Karoline Freeman, and Jackie Walton added to study investigators, and affiliations added. Table 10 added to data requested. Table 10 lists the data items which would improve the analysis, but are not yet available for linkage into the dataset (but may soon be). Work package 2 added that different definitions of round length (beyond 2 vs 3 years) will be investigated if time permits. Timings updated on gantt chart. Other minor corrections to wording.

Version 1.3 Minor additional amendments requested by Public Health England. Expanded details of cause and date of death, and sources. Expanded explanation of cancer type outcome to clarify it covers screen detected only, and symptomatic and screen detected combined. Clarified that reader threshold analysis will have to consider number of readers. Provided clarification of date of data extract. Tables renumbered to fit PHE preferred structure.

Version 1.4 Clarification of outcomes to match clinicaltrials.gov registration, including: clarification that outcomes labelled 'health outcomes' (not intermediate outcomes) are the primary outcomes, clarification that breast cancer mortality is the first outcome (due to greater power to detect differences than all-cause mortality), moving false positive recall from intermediate to primary outcomes due to its known association with anxiety, and reordering of presentation to match clinicaltrials.gov entry. More detail added to measurement of characteristics of cancer detected. Outcomes, whilst reordered in presentation remain unchanged, Addition of preliminary analyses for test threshold to determine whether instrumental variable assumptions are met and therefore whether this analysis is possible. Clinicaltrials.gov identifier added.

Summary of Research

Background

There is a lot of debate about whether the benefits of breast cancer screening outweigh the harms. The UK Independent Review (1) led by Sir Michael Marmot estimated that breast cancer screening in the UK saves 1300 women's lives every year; however 70,000 women each year are unnecessarily made anxious after the screening test (mammography) shows potential signs of cancer, but which are found to be benign (false positive results). Another 4000 women are given unnecessary cancer treatment, because they have a cancer detected at screening that is so slow growing that it would never have harmed them or given them any symptoms in their life (overdiagnosis).(1) These are the best national estimates we have, but Sir Michael Marmot and colleagues based their estimates on randomised controlled trials from the 1970s, and did not have sufficient data to update them to more recent tests or treatments. Current tests use much higher resolution so the radiologists can identify smaller,

earlier stage changes, and modern treatment is much more effective. This affects how much benefit screening provides, so the current balance of benefits and harms are uncertain.

International Variation in Breast Screening

Different countries offer diverse versions of breast screening, because it is uncertain which is best. In the UK we offer screening every three years, which is the longest interval between breast screens in the world. In the US screening is undertaken every 1-2 years, and in most of Europe every two years. In practice in England there is a lot of variability in round length received by individual women, due to the policy and procedures, and workload of individual breast screening centres. In England overall, 4% of women are recalled for further tests because their mammograms show suspicious signs, with other countries recalling as few as 2% (Denmark) or as many as 10% (USA). However the recall rate varies greatly between centres, because there is uncertainty about what is the best level.

Our Approach

We need to understand which version of breast screening offers the most benefit with the least harm. The best evidence for this would be from a Randomised Controlled Trial (RCT). However, such trials are very expensive, often do not include sufficient women, and need at least 10 years follow-up. So a trial that started recruiting now would not give answers until the year 2036. Instead, we suggest making use of the data available from offering different versions of breast screening to over 13 million women in England over the last 25 years, as a retrospective observational research study. We have carefully designed the design and analysis using methods that allow causal inference to be made from observational data.

The Research Plan

In this very large observational study we will analyse the records of women screened in England between 1990 and 2018, with follow up available as to whether they got breast cancer (from the English Cancer Registry) and whether they died (from the civil registration mortality file). We will analyse the ages and frequencies women are invited and the proportion of women we recall for further tests affects the benefits and harms of screening. These benefits and harms will include numbers of false positive recalls, overdiagnosed cases, and mortality. We will also analyse the mechanism of action for any changes, how changes to number and nature (eg histological grade and type, stage and size) of cancers detected at screening affect number and nature of cancer detected symptomatically in the years after screening, numbers overdiagnosed and life years saved. We will explore how detecting greater numbers of Ductal Carcinoma in Situ (DCIS, i.e. most common form of potential precursor of invasive breast cancer) affects the benefits and harms of screening.

Using Findings to Change NHS Practice

We will use the findings to inform the UK National Screening Committee and revision of the English quality assurance guidelines for breast cancer screening. This will depend on results, but could for example include revising the targets for proportion of women recalled, of DCIS detection rates, or level of flexibility in the screening interval target.

1. Background and Rationale

<u>What is the problem being addressed?</u>

We do not currently know which version of breast cancer screening gives most benefit and least harm to the women who are screened. In particular, how often should women be

invited to breast cancer screening? What is the appropriate age range to offer breast cancer screening? Which recall threshold should we use for the test? Which types of mammographic abnormality should be investigated further?

There is huge national and international variation in breast screening practice. This is due to the lack of evidence of which version gives most benefit and causes least harm. In England breast screening is offered every three years to women aged 50-70 using two-view mammography (2 x-rays) of each breast. The decision to use a three yearly interval was made in 1985.(2) Other countries screen either every year (US(3)) or two years (most European countries). The upper age limit was extended in the UK from 64 to 70 years in the NHS Cancer Plan in 2000(4) on the basis of pilot studies of acceptability and uptake only.(4) There is an ongoing trial of extending breast screening to between ages 47 and 73 (first results to be reported in 2026)(5) but there has been no evaluation of the previous age extension to 70 years.

In England, of women attending breast cancer screening, 3.9% are recalled for further tests from each screening appointment(6), compared to 2.5% (Denmark(7)) or 10.6% (USA(8)). The percentage of women recalled is a marker for the radiologists 'recall threshold' which the radiologist can change by electing whether or not to recall moderately suspicious findings by type, or even by changing the workstation/workflow.(9) We need to understand the impact of these different versions of screening not only on proxy outcomes such as numbers of breast cancers detected, but on outcomes which affect the women screened such as: cancer and treatment-related morbidity; mortality; and overdiagnosis and associated overtreatment of breast cancer which would never become symptomatic in the woman's lifetime.

There is also uncertainty about which types of mammographic abnormalities and cancers we should aim to detect at screening. Since the trials of breast screening in the 1970s screening technology has improved markedly in resolution. Now we can detect smaller cancers, and smaller microcalcifications that may be associated with DCIS. Similarly, we detect more of these when we reduce the recall threshold. We do not know what the balance of benefits and harms are of detecting these very small cancers and 'pre-cancer' abnormalities, they may increase the life years saved or the number of women harmed by overdiagnosis.

The ideal study design to answer specific questions such as this would be the randomised controlled trial (RCT). However this approach is impractical given the multiplicity of unknowns, and is not financially feasible for most changes to screening. However, given the natural variations within the records of the English national screening programme, a large observational design study is possible, enabling sufficient power and follow-up to real clinical outcomes.

Why is this research important in terms of improving the health and/or wellbeing of the public and/or to patients and health and care services?

Over two million women attend breast cancer screening each year in England. There is an ongoing international debate about the balance of benefits and harms. The UK Independent Review chaired by Sir Michael Marmot concluded that breast screening saves 1300 lives each year in the UK; however results in overdiagnosis and unnecessary treatment of another

4000 women, and 70,000 women receive false positive results and are subject to the associated anxiety.(1) This review was based on RCT evidence from the 1970s. These trials did not provide sufficient evidence to optimise screening interval, recall threshold, or age of eligibility which has led to the international variability. In addition, since these RCTs breast screening technology has improved and many smaller features can be detected, in particular microcalcifications associated with DCIS.

Since its inception in 1988, 10 million women have attended breast screening, receiving a range of different versions of screening. (e.g. invited at different ages with different intervals between their screens etc.) We have, nationally, kept excellent records regarding the versions of screening each woman was offered, whether and when they developed cancer, and whether and when they died. Women have received different screening strategies providing a natural experiment. The screening strategy received is driven by centre level variables rather than the woman's individual characteristics and prognostic factors, reducing potential bias due to confounding. We propose using these data to understand which versions of screening are most effective, and determine which is the best version to offer to the two million women who attend each year going forwards. This would mean that breast screening can be standardised to the version which gives the most mortality and morbidity benefit and least overdiagnosis and false positive harm. We will focus equal attention on the possibility of more screening or less screening.

This research will also provide evidence to minimise the variability in care that women in England currently receive. Some breast radiologists recall 18% of women whilst others just 2%(9). There is huge inter-centre variability in number of cases of DCIS detected, due to differing beliefs in the benefits and harms of detection. Clear evidence linking to benefits and harms would drive reduction in variability. Our PPI team have identified this as a priority.

This research will impact the health of those attending breast cancer screening through working with the UK National Screening Committee if our research suggests a substantial programme modification, and Public Health England (PHE) to modify the breast screening quality assurance guidelines, linking targets to maximising benefits and minimising harms of screening.

Finally, this research also has the potential to influence future policy decisions. Our proposed analysis of mechanism of action will enhance our understanding of the benefits and harms of detecting different cancer and precancer types at screening. This will be important when making future policy decisions about new breast screening testing technology (such as incorporating fast MRI, tomosynthesis or artificial intelligence readers into the screening programme), as they all detect a different spectrum of disease to current screening. Our research will provide some evidence to link spectrum of disease to benefits and harms of screening.

Review of existing evidence - How does the existing literature support this proposal?

In 2016 the US Preventive Services Task Force undertook a comprehensive evidence review of the frequency and ages women should be offered breast screening.(10) They recommended mammography every two years for women aged 50 to 74 years. The review conclusions for screening frequency were based on two observational studies including 941,938 women from the US.(11, 12) These studies were not able to account for

confounding, as they did not have information on which screening intervals were shortened for clinical reasons. We have this information in the English dataset, along with the exact round length in days. Significantly, the US review recommended future research of the type we are proposing here, for example examining at what age screening should be discontinued, and outcomes from DCIS detection.

The last review concerning breast cancer screening commissioned by the UK National Screening Committee was undertaken by our research group in 2019.(13) This review examined whether to make changes to the screening test. This review highlighted that additional testing could detect extra cancers, but they tended to be smaller, node-negative, lower grade invasive cancers (which have a good prognosis). These are similar to the extra types of cancer detected when reducing the recall threshold for mammography(14, 15). The review highlighted the research needed to determine whether these would be beneficial in terms of reduction of mortality or morbidity, or would represent overdiagnosis.

This proposed study would be an order of magnitude larger than the studies cited in the UK and US reviews, giving sufficient statistical power to investigate important outcomes for the woman screened, rather than proxies. We have individual patient data available, with very low rates of missing data and loss to follow-up as a result of the standardised software and mandatory reporting mechanisms in the English Breast Screening Programme.

There are previous UK studies which have investigated impacts of recall threshold. Blanks and colleagues investigated the relationship between needle biopsy rate (proportion of women screened who are recalled for further tests and one of those tests is a needle biopsy) and cancer detection rate. (16) They found a positive correlation, with diminishing returns at higher rates of biopsy. They did not present data on interval cancers or health outcomes. In a similar study Blanks and colleagues investigated the relationship between English and Dutch breast screening centres recall rates and cancer detection rates.(14) They concluded that increasing recall rates was associated with increasing detection rates of DCIS, although this relationship may have been subject to Country level confounding as Dutch recall rates were systematically lower. They did not present data on interval cancers. Burnside and colleagues found an inverse relationship between recall rate (a proxy for test threshold) and rate of interval cancers at UK breast screening centres, equivalent to one fewer interval cancer for every 80-84 recalls.(17) Duffy et al. found an inverse correlation between detection of DCIS and subsequent interval cancers. (18) All four of these studies were at the centre level rather than the individual or reader levels. Our proposed analysis uses data at the individual, reader and centre level so there are more units of analysis and we are able to adjust for individual or reader level variables. Further, this allows us to design the study explicitly considering whether we can make causal inference, whereas previous studies simply reported correlations. Finally, none of these previous studies extended analyses beyond cancer detection or interval cancer development, to longer term health outcomes, which we propose to do.

There was one randomised controlled trial comparing annual to three-yearly screening in the UK. They found no statistically significant difference between the two arms in predicted survival using Nottingham Prognostic Index, but the confidence intervals were wide and it was underpowered.(19) There is limited evidence from observational studies examining different screening frequencies.

The previous extension of the upper age limit of screening from 64 to 70 is supported by women of these ages being included in some of the original RCTs of screening, although the balance of benefits and harms by age has mainly been assessed by economic modelling relying heavily on assumptions.(21) Screening has a higher cancer yield in older women, but

also increased risk of overdiagnosis.(21) In a recent survey of 21 countries, all of them screened up to at least age 70.(22) Recent advances in causal inference methods for observational data have been applied to examining the benefits and harms of colorectal cancer screening (20) but we believe we are the first group to propose to apply this approach to breast cancer screening.

4. Aims and Objectives

How does age of eligibility, screening interval and recall threshold affect the benefits and harms of breast cancer screening?

Aims:

1. To understand how age of eligibility, screening interval, and recall threshold for breast cancer screening affect benefits and harms including false positive recalls, overdiagnosis and mortality.
2. To inform revision of the quality assurance guidelines for breast screening centres based on maximising benefit and minimising harm from breast screening

Work Package 1: Database development and access

Objective 1: To assemble, clean and assess the quality of the combined datasets

a. To obtain approvals to re-use the observational dataset of 13 million women offered breast screening
b. To clean data and describe quality

Work Package 2: Causal links between age of eligibility, screening interval, recall threshold and health outcomes

Objective 2: To analyse the causal effect of age of eligibility, screening interval, and recall threshold on intermediate outcomes (numbers of breast cancers detected at screening by cancer type, interval cancers, false positive recalls) and health outcomes (mortality, morbidity, overdiagnosis).

Work Package 3: Pathway to impact

Objective 3: To apply findings to inform changes to practice, including changes to the NHS Breast Screening Programme consolidated standards.

5. Research Plan/Methods

The project team brings together direct experience of data linkage and analysis of the database (Taylor-Phillips/Clarke/Wallis/Kearins) with internationally renowned analytical expertise related to medical tests (Deeks/Sitch) and causal inference from observational data (Sterne), observational data quality (Brettschneider), and clinical expertise in breast radiology (Wallis/Given-Wilson/Wilkinson) and breast pathology (Pinder). To optimise local implementation of findings members/Chairs of English national decision-making groups are involved (Given-Wilson/Wilkinson/Taylor-Phillips/Wallis), and the Public Heath England national Quality Assurance lead (Kearins), along with implementation science expertise (Currie). We work in partnership with Independent Cancer Patients Voices (Gath/Radin/Walker) to include the patient voice at every stage. All co-applicants will be involved in every work package (and will contribute to study development meetings), the lead centre is shown in brackets for each.

Work Package 1: Database development and Access (Led by Olive Kearins, PHE Screening and Sian Taylor-Phillips/Julia Brettschneider, Warwick)

Main data set (objective 1a): Our research team have already linked breast screening data from 80 English centres with English Cancer Registry and civil registration s mortality file data for 10 million women attending 35 million breast screening appointments, between 1988 and 2018 (Taylor-Phillips, POSTBOx, NIHR, 572k). Data were extracted from centres up to 2018, but a date cut-off of date of first offered appointment up to end Dec 2016 was implemented in the previous data transfer to Warwick (for the previous POSTBOx project). The dataset for ATHENA-M will be based on this same database, but with the following additional data transferred to Warwick for analysis i. An additional 3 million women, who were offered screening but never attended, to enable intention to treat analysis ii. an additional outcome of cause of death, in particular breast cancer death. iii. Updated linkage to more recent follow-up in cancer registry to end 2018 and civil death registration to date of linkage in 2021. This transfer will be in two parts. Firstly, the expanded dataset with the extra 3 million women, without updating linkages or adding the extra outcome of cause of death. This first dataset will ensure there are data available in the event of issues arising in the reorganisation of Public Health England. The second transfer will contain updated linkages to the cancer registry (to end 2018/2019 as available) and civil death registration (to 2021), and extra outcomes of cause of death. Further updates to the database based on later extracts from NBSS (post 2016) and updated linkage to civil death registration, cancer registry and/or BSselect may be made to ensure these data are as current as possible. Summary of requested tables is in appendix 2, and full definition of all variables is in a separate document. Inclusion criteria is in appendix 3. This cohort definition gives 2.3 million women invited never screened between 1988 and 2016 when aged 47 to 73. These are additional to the 10.5 million women with at least one screened episode between 1988 and 2016 when aged 47 to 73.

These data are more complete, and validated than any other breast screening data worldwide. They include >35 million screening appointments recalling >2.2 million women, detecting >250,000 cancers, with >100 million person-years of follow up to >500,000 subsequent deaths. Analyses are well powered, (sample size calculations are given in work package 2). There are little missing data, for example the radiologist's decision is missing in 4,527 (0.01% or 1 in 10,000) episodes, reader identifier is missing in 7,571 (0.02%) episodes. Many of these cases of missing data are from the early years of screening, between 1988 and 1993). Data are missing for the outcome of screening (cancer detected or not) in 6,992 (0.02%) episodes. Many of these are due to women dying between being recalled for further tests and attending for those tests, for these cases we have mortality data (which is the outcome in some analyses, and used for censoring in other analyses), so this may be an overestimation of data missingness.

The variables in the current dataset include:

- Details of each screening appointment (pseudonymised centre identifier, pseudonymised woman identifier, date (available for every appointment for each woman so screening interval is known), whether she attended, the decision of whether to recall and the pseudonymised identity of the first reader examining the mammograms, the decision of whether to recall and the pseudonymised identity of the second reader examining the mammograms, the decision and pseudonymised identity of the arbitrator of the decision whether to recall, whether she was recalled for further tests, whether she had a biopsy, whether she had cancer detected at screening (from biopsy results), whether this episode was part of a trial eg AgeX)

- Details of the woman (identifier to link to every screening episode she was invited to between 1988 and 2018, month and year of birth, index of multiple derivation from postcode)
- Details of screen detected cancers obtained through linkage to the English Cancer Registry (Histological grade, tumour node metastasis (TNM) stage, size, hormonal status (oestrogen receptor (ER),progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2)), nodal involvement, treatment received (breast surgery, axillary surgery, radiotherapy, chemotherapy))
- Details of symptomatically detected cancers obtained from the English Cancer Registry, including interval cancers detected between screening rounds and cancers detected in the years after screening (date of diagnosis, histological grade, TNM stage, size, hormonal status (ER,PR,HER2), nodal involvement, treatment received (breast surgery, axillary surgery, radiotherapy, chemotherapy)
- Date of death (from the civil registrations mortality file, which is populated with data from the Office of National Statistics)

We have existing permissions (NHS ethical, office of data release (ODR)) to use the data set to analyse the effect of recall threshold on rates of overdiagnosis only. We have already combined screening data from all 80 English centres within PHE, completed linkage to the Cancer Registry, and MBIS. Full permissions have been granted for transfer of these data to the University of Warwick, a contract is in place for the transfer, and transfer of these data is complete

For the ATHENA-M project we will require the following. Firstly, updated permissions from ODR and NHSREC to include: extending the analyses to the frequency and age of eligibility research questions, to include outcomes beyond overdiagnosis, to extend the cohort to women never screened, and to extend the groups authorised to hold these data to include Birmingham and Bristol (which will require site specific data security plans). There is a small amount of extra data linkage within PHE Birmingham using the same methods as for the previous project, to extend the cohort to women never screened and add the outcome of cause of death to calculate breast cancer mortality. The data linkage work itself will take less than a week as we have pre-existing code from previous projects. It will not require any additional data to be extracted from breast screening centres, this was all completed as part of the previous project. We have conditional permissions from the Breast Research Advisory Committee for ATHENA-M. Approval from this group automatically starts the ODR permissions process, so we expect to be close to achieving these permissions by the project start date. The ATHENA-M project answers research questions of importance to Public Health England, and identified by Public Health England, and so is covered by the existing section 251 approvals, and does not require independent section 251 approval. HRA approval has been granted (21/LO/0120).

Data quality and Cleaning (objective 1b): Datasets will be cleaned. Quality assessment including data completeness for each item, test for missingness at random, and accuracy vs other validated sources will be reported. Other validated sources will include the KC62 annual Korner returns and the Association of Breast Surgeons (ABS) audit. Changes over time, and by centre will also be reported. We will also consider potential misclassification, such as the possibility that mortality is more likely attributed to breast cancer in screened women. This will be published in a journal article, and internal PHE report.

Work Package 2: Causal links between Age of eligibility, Screening interval, Recall Threshold and health outcomes

We will define each question by specifying the eligible patients, experimental and comparator interventions and outcome for a 'target trial' whose results we aim to estimate using the observational data.(23) This target trial approach proposed by Hernan and colleagues(23) gives a framework to avoid potential biases that would prevent causal inference, such as immortal time bias (where the outcome cannot occur during part of the follow-up).

Our comparisons are between different versions of invitations to screening. We have carefully considered time varying confounding in our analysis plans. There is a potential for time varying confounding if prognostic factors for outcomes of interest influence women moving from one treatment group to another over time, and if follow-up for individual women is split between treatment groups. Our proposed analyses explicitly account for time-varying confounding and use appropriate methods to address it. Importantly, in analyses assessing the effect of extra invitations to screening or different screening intervals it is unlikely that individual women's prognostic factors directly affect receipt of invitations (they would by contrast affect uptake of invitations, which we do not propose to assess). This is because each centre sends invitations to all eligible women regardless of their personal characteristics. However, we will examine whether centre level characteristics that may be associated with prognostic factors for outcomes of interest also affect invitations to screening. We will consider a range of centre level characteristics for potential inclusion, such as index of multiple deprivation of population served, arbitration system used, centre size, quality assurance region, and quality assurance indicators.

We will use survival analyses approaches allow for different lengths of follow-up, and will adjust standard errors account for clustering by centre,reader/reader pair, and/or screening batch where appropriate. Breast cancer treatment has developed significantly and associated mortality has decreased. We will therefore adjust for calendar time, using smoothing splines, in all time-to-event analyses.

Screening programmes can exacerbate inequality by lower uptake in lower socioeconomic groups. In addition to the main analysis we will evaluate whether the outcomes from the evaluated changes to screening age of eligibility, interval and threshold differ by index of multiple deprivation (a proxy for socioeconomic status derived from the postcode of the woman's most recent address). We will present results for how each change affects overall outcomes, and by groups according to index of multiple deprivation.

Measuring Outcomes:

Intermediate outcomes

1. Interval cancers detected within 3 years of screening. An interval cancer is a cancer detected symptomatically in the interval between breast screening appointments. These are all biopsy proven, with records taken from the English Cancer Registry. It is a binary outcome for each episode of screening for each woman. We will use the NBSS data to exclude screen detected cancers from this measure. Interval cancers has been identified by the UK National Screening Committee, Public Health England and our PPI team as an important intermediate outcome because reducing interval cancers is not associated with increasing overdiagnosis in the same manner as number of cancers detected.
2. Cancers detected at screening, taken from the results of screening recorded by the Breast Screening Programme in the National Breast Screening Service (NBSS)

database. It is a binary outcome for each episode of screening for each woman. These are all biopsy proven as per screening programme standards. Definition includes any invasive cancer or Ductal Carcinoma in situ (DCIS) or Lobular Carcinoma in Situ (LCIS) of the breast, using standard definition of cancer registry and screening programme) with subgroup with invasive cancer only also reported. Records are complete because there are quality assurance mechanisms to ensure complete recording of these as part of the standard KC62 reporting procedures.

3. Screen detected cancer characteristics. This is defined by variables from the English Cancer Registry, including histological grade, cancer stage, and cancer size. These intermediate outcomes were chosen because the proposed mechanism of action of breast screening benefit is through detection of smaller earlier stage cancers. This outcome can be difficult to interpret as increased detection may be associated with mortality benefit or overdiagnosis harm. This will be measured for screen detected cancers only, (and for a combination of screen and symptomatic cancers, where differences between exposed and unexposed represents the stage shift of exposure: see stage shift outcome). [In detail TNM stage (size, nodes, distant metastases), alternative nodal count to match Z11 and POSNOC trials 1/1-2/3+, DCIS vs invasive, invasive grade 1/2/3, DCIS grade 1/2/3, DCIS surgical size (mm), Invasive surgical size (mm), invasive cancer type, hormonal status, Nottingham Prognostic Index].

Health Outcomes or close proxies (Primary outcomes)

1. Breast cancer mortality, as defined by the English Cancer Registry (where the initial source is the same but linkage to NBSS data more complete as uses multiple fields rather than just NHS number) checked against cause of death from the civil registrations mortality file. These sources have limited death data pre-1997, so we will also extract deaths data from NBSS to determine whether it can provide adequate quality, and if so use pre-1997. It is a binary outcome for each woman, with different lengths of follow up. Reported as cumulative incidence over all follow-up time, with focus on 10 year and 13 year to match previous systematic reviews.
2. All-cause mortality for everyone in the included cohort. This is taken from the civil registrations mortality file, which is populated by the Office of National Statistics. It is a binary outcome for each woman, with different lengths of follow up. Reported as cumulative incidence over all follow-up time, with focus on 10 year and 13 year to match breast cancer mortality.
3. Overdiagnosis. Overdiagnosis will be inferred as any difference between the study groups in total cumulative incidence of cancer (screen and symptomatic detected) after sufficient follow-up (the compensatory drop method). Therefore the outcome measured will be difference between groups in sum of breast cancers detected at screening and symptomatically. We will carefully consider length of follow-up in interpreting results relating to overdiagnosis, as insufficient follow up results in overestimation. We will publish a full protocol detailing all analysis methods and outcomes before commencing analysis to prevent selective reporting of outcomes or other analysis elements. Reported as cumulative incidence over all follow-up time, with focus on 10 year and 13 year to match breast cancer mortality.
4. Stage shift. This is defined by variables from the English Cancer Registry, including histological grade, cancer stage, and cancer size. The proposed mechanism of action of breast screening benefit is through detection of smaller earlier stage cancers. This will be measured for a combination of screen and symptomatic cancers, where differences between exposed and unexposed represents the stage shift of exposure. [In detail TNM stage (size, nodes, distant metastases), alternative nodal count to match Z11 and POSNOC trials 1/1-2/3+, DCIS vs invasive, invasive

grade 1/2/3, DCIS grade 1/2/3, DCIS surgical size (mm), Invasive surgical size (mm), invasive cancer type, hormonal status, Nottingham Prognostic Index].

5. Morbidity. Here we are interested in the morbidity associated with breast cancer treatment. We measure this as four outcomes: women receiving breast surgery, axillary surgery, radiotherapy, and chemotherapy within one year of diagnosis. Each of these are binary outcomes for each episode of cancer for each woman.
6. False positive recalls. These are women recalled from screening for extra tests but those extra tests do not indicate cancer. These data are taken from the NBSS computer system which automatically records who is recalled. This is measured as any women who were recalled but did not have cancer detected in follow up tests. It is a binary outcome for each episode of screening for each woman. This is an important harm of screening because it is associated with increased anxiety in women screened.

We will investigate the mechanism of action linking the three exposures to the outcomes. In particular we will investigate the relationship between characteristics of cancer detected, and women's outcomes such as overdiagnosis and mortality and morbidity associated with treatment. This analysis has dual purposes: firstly when inferring causation from observational research, in addition to appropriate accounting for confounding, it is important to elucidate the mechanism of action. Secondly, it is an important research output in its own right, to provide the evidence base for policy-makers to link the characteristics of cancers detected to benefits and harms of screening, when assessing a range of changes to screening. In the UK and Australia this is referred to as the linked evidence approach, and the US Preventative Services Task Force refer to it as the dotted line in the analytic framework.

How do two extra invitations to screening between the ages of 65 and <71 years affect all-cause and breast cancer specific mortality, overdiagnosis, treatment associated morbidity, and false positive recalls, in women already invited to screening age 50-64? (analysis led by Jonathan Sterne, carried out at Bristol)

The NHS Cancer plan(4) extended the age range of breast screening in England from 50-64 years to 50-70 years, which was rolled out across all 80 centres between the years 2000 and 2006. The first sites to adopt the age extension were those piloting the introduction of the assistant practitioner role to assist radiographers in their work, as part of the introduction of the four tier workforce.(24) After these pilot sites, timing of roll out to other sites was dependent on funding, staffing and extra mammography equipment in place. The provision of extra equipment was lottery funded.

Within the 2000-2006 time period we will compare outcomes in women who were and were not offered two extra rounds of screening, adjusting for temporal, centre and individual level confounding. We will use the approach described by García-Albéniz et al.,(20) to emulate a weekly series of trials in which eligible women who have not yet been offered additional screening are assigned to receive or not receive screening in the coming month. Women will become eligible for the first trial on their 64th birthday*, and each woman will contribute to subsequent trials providing that she remains eligible until the day before she turns 71. We will make two sets of comparison: (a) no additional screening compared with one or more invitations to additional screening; and (b) at least two additional invitations compared with no or one additional invitation. For each weekly trial, follow-up in women in the "less screening" group will be censored at the time that receive sufficiently many invitations for inclusion in the "more screening" group. We will use inverse probability weights to adjust for

selection bias introduced by such censoring. Standard errors will be adjusted for inclusion of women in multiple weekly trials. Analyses will be by intention to screen, with women included in each group regardless of whether they attended screening. We will limit the analysis to years in which some centres had and others had not rolled out the screening age extension. We will model changes in the rate of each outcome with calendar time using cubic splines. Self-referral to screening in the years after screening will attenuate any effect, but fewer than 12% of women self-refer. **Preliminary investigation will be necessary to ascertain how the upper age limit of 64 was implemented, as it may in practice have been until a womans 65th birthday*

How does screening intervals of between 15 and 27 months, compared with intervals of between 28 and 40 months affect cancer stage at diagnosis, treatment associated morbidity, and false positive recalls in women attending breast cancer screening age 50-70 years? (analysis led by Jonathan Sterne, carried out at Bristol)

We will use the 'clone, censor and weight' approach described by Hernán.(25) Women will be eligible if screened between ages 50-70 years, and follow-up will start 15 months after the date of first invitation to screening. Follow-up for each woman will be duplicated ('cloned'), with one copy assigned to each treatment strategy. Follow-up for each cloned copy will be until women deviate from the strategy assigned that copy, because they are invited to screening too late (in the more frequent screening group) or too early (in the less frequent screening group). We will model the probability of being screened over time, in order to derive inverse-probability of screening weights. Moving house (and screening centre) will be included as a covariate, as this can prompt a shorter screening interval. These weights will be used to adjust for the selection bias introduced by censoring follow-up at the time of deviation from assigned treatment strategy. Planned shorter screening intervals due to suspicious findings at baseline will be excluded.

Additional analyses of will be conducted to investigate whether the effects of screening interval on health outcomes vary according to age group at screening (50-60 or 61-70 years).

Different definitions of round length will be investigated if time permits. This may include women attending family history screening and/or women whose round length was delayed by the COVID-19 pandemic.

How does recall threshold affect all-cause and breast cancer specific mortality, overdiagnosis, treatment associated morbidity, and false positive recalls in women attending breast cancer screening age 50-70 years? (analysis led by Jon Deeks and Sian Taylor-Phillips, undertaken at Birmingham and Warwick [Taylor-Phillips and Freeman working across both universities])

We will estimate the effect of screening threshold, using instrumental variable approaches or other analyses with consideration of causal inference. The exposure or instrumental variable will be the rate of recall of the previous 5000 cases screened by the reader (or reader pair). The readers' recall rate for previous cases may be an appropriate instrumental variable to because its effect on the outcome (for the current case) is only via the readers' threshold for the current case. This assumption will be evaluated. Instrumental variable definitions will be finalised in preliminary analyses that aim to maximise their association with recall probability, in datasets from which all outcomes have been removed. At this stage we will consider different numbers of previous cases to inform the instrumental variable definition. 5000 represents the mandated minimum per year for each reader, according to English quality assurance guidelines(26) and would give a reasonably precise estimate of recall threshold (if

the reader recalled 4% of cases the 95% binomial exact CI is 3.5% to 4.6%). We know there is substantial inter-reader variability in recall threshold, and intra-reader temporal changes (changes within a single reader over time) are much smaller, justifying increasing the number of previous cases to reduce statistical variability in the instrumental variable. We will conduct sensitivity analyses to assess whether results are robust to changes in the instrumental variable definition. We have data for a full range of thresholds for every year since 1988, and therefore have sufficient data to adjust for temporal changes. We will model changes in rates of outcome with calendar time using cubic splines.

Three models will be developed for three time horizons (shown in brackets): 1: The effect of threshold at the final screening round (27 years) with adjustment for previous attendance record, so follow-up is not contaminated with screening invitations. 2: The effect of threshold in a screening round (3 years), adjusting for clustering of screening episodes within women, the women's age at screening, and previous attendance patterns, with outcomes limited to false positive recalls, interval cancers, and stage of cancer detection. 3: If feasible, the cumulative effect of threshold over all screening rounds including the woman's entire screening history and follow-up post screening (27 years)

Modern mammography screening in England uses two readers and arbitration of discordant assessments. Any subsequent intervention to change the recall threshold would most likely act upon each individual reader. We will use modelling approaches to evaluate the impact of test threshold on screening outcomes; we will investigate the impact of changing the process (adjusting the threshold of both readers), changing the threshold for an individual reader, the impact of one vs two readers, changing the threshold for one or both, and the pairing of readers. We will use instrumental variables to describe the recall threshold of each reader, in these proposed models. We will also model the overall effect of changing recall threshold, using a predictor for the single readers decision in the older cases where there was only one reader, and the reader combination's combined threshold in later years. This will show the relationship between the recall threshold of the system and women's outcomes.

We will investigate the mechanism of action of any effects, through intermediate outcomes. For example we expect that reduced recall threshold may increase cancer detection but change the spectrum of disease identified towards more small cancer detection (possibly of low histological grade) and also DCIS, which may in turn affect mortality and overdiagnosis. This will help our understanding of what we should aim to detect at breast screening.

A further exploratory analysis of instrumental variable approaches to recall threshold specifically for DCIS will be undertaken, as DCIS is detected predominantly through microcalcifications, and there is vigorous debate about the benefits and harms of detecting DCIS.

We will also investigate the variability between centres and readers, and its implications for screening QA targets.

Additional analyses of will be conducted to investigate whether the effects of recall threshold on health outcomes vary according to age group at screening (50-60 or 61-70 years).

Sample Size Requirements

This observational data analysis does not require a formal sample size calculation. We will include every eligible case within the available cohort in analyses. However, we have included some calculations here to ensure that there are sufficient available data to enable meaningful analysis. Previous analyses have shown clustering of the proposed outcomes within centres and readers is negligible (ICC<<0.0001), hence the cluster of observations is

not accounted for in these calculation; however, we will adjust for these in the analyses where appropriate.

Table 1. Sample size requirements for detecting differences within the dataset. All calculations are for 90% power at 5% significance level. There are more than 30 million screening appointments in the dataset, from more than 10 million women.

Outcome	Baseline value	Change in value	Number required in each group to detect
Intermediate outcomes			
Recall rate	3.9%	0.1%	777,920
Cancer detection rate at screening	8.4/1000	0.5/1000	679,499
Small invasive cancer detection rate at screening	3.4/1000	0.3/1000	756,401
Interval cancer over 3 years	2.9/1000	0.3/1000	640,353
False positive recalls	3.1%	0.1%	641,114
Health outcomes			
All-cause mortality	0.6050 per 100 person-years	0.6025 per 100 person years	804,400
Breast Cancer mortality	0.0443 per 100 person years	0.0433 per 100 person years	880,116
Overdiagnosis	1.3% of women invited for screening for 20 years(1)	0.1%	279,869

We will deliver three journal articles, one for each of the three questions investigated: age of eligibility; recall threshold and round length. We will submit a fourth journal article exploring the link between detection of different cancer types and outcomes, with a focus on DCIS in particular. We will also present results to the UK National Screening Committee, and inform redrafting of the English national quality assurance standards for breast screening (see work package 3).

Work Package 3: Pathway to impact (Led by Olive Kearins, PHE and Sian Taylor-Phillips, Warwick)

In addition to journal publication there are three strands to our dissemination strategy. 1. Influencing national policy directly via the UK National Screening Committee 2. Influencing national practice through national professional guidance and communication to health professionals 3. Communication to the public

<u>1. Influencing national policy directly via the UK National Screening Committee (UKNSC)</u>

The UK National Screening Committee is responsible for all national screening policy, including making major changes to screening programmes such as age of eligibility or frequency of screening. We will present our results in person to the UK National Screening Committee Adult Reference Group (ARG) which is responsible for all adult screening programmes (on which Professor Taylor-Phillips sits). If our results suggest that a major change is appropriate then we will submit a formal request via the annual call or directly to the ARG, and the UKNSC would then undertake a systematic review putting our research into context and make a national decision on that basis. These decisions are directly

nationally implemented with associated budget so encounter fewer barriers to implementation than guidelines.

We will provide evidence on how the changes would affect the benefits and harms of screening overall, and on how the changes would differentially affect women of different socioeconomic status (using our analysis of a proxy for this, Index of Multiple Deprivation). This will allow the UK NSC to consider the implications of any decision on inequalities, as required by UKNSC criterion 12. (There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public).

The UK National Screening Committee will be involved throughout the process, with Professor Bob Steele (UKNSC chair) and Professor Anne Mackie (Director of PHE Screening) joining our national policy advisory group, Dr Ros Given-Wilson (Chair of ARG) a co-applicant, and regular written updates given to key personnel such as John Marshall (PHE Screening evidence lead). Letters of support from Professor Steele and Professor Mackie are attached (appendix 1)

2. Influencing national practice through national professional guidance and communication to health professionals

If our research were to suggest that there is a need to reduce variability in screening intervals, or change test threshold this would be implemented through national professional guidance, changes to the breast screening programme specific operating model,(27) and influencing changes to practice. Here Professor Graeme Currie (Professor of Implementation Science, co-applicant) is advising on strategies to maximise implementation. Our approach will be:

i. To engage in the process of redrafting the national professional guidance,
ii. To identify and engage national champions for practice change who will drive change through their networks.

Redrafting the national professional guidance: Our findings will be used to inform the regular update of the NHS Breast Screening Programme Consolidated standards, and the Professional Guidance for Breast Cancer Screening Radiology (objective 3, led by Olive Kearins, Breast Screening Research & Data Lead, PHE Screening). These are used by commissioners to develop the service specifications when commissioning breast cancer screening, by the screening QA service to inform quality interventions and by radiologists in assessing their own performance. We will use the findings from objectives 1 to 3 to propose changes to quality metrics to align with maximising benefit and minimising harm, cognisant of effects on inequalities (via the Index of multiple deprivation proxy), and of statistical variability and the need to define targets measurable every year at each centre. Our analysis of threshold will inform consideration of targets for recall rate (currently <5%), invasive cancer detection rate (currently ≥5.7/1000), small cancer detection rate (currently≥3.1/1000), DCIS detection rate (currently ≥0.6/1000, all for previous attenders), and round length (currently 3 years). (26, 28) PHE are implementing a live data monitoring system, so centre-level performance towards QA targets can be measured. PHE will consult widely in developing revised guidance and standards, engaging a wider range of practitioners to identify practical issues, misunderstandings, attitudes, and context from a wider range of stakeholders including radiologists who are less research-involved, and radiographers, pathologists, breast clinicians, breast care nurses, administrative staff and centre managers. Whilst the guideline development is led by Public Health England, implementation through commissioning is the responsibility of NHS England. Jeff Featherstone, Head of Public

Health Commissioning and Operations for NHS England and NHS Improvement, and Cath Fenton (regional NHS England) will sit on our advisory group. They will advise early throughout the study on how practice may be influenced through commissioning, including barriers and enablers.

Identify and engage national champions for practice change: Our strategy focuses on a no-surprises approach, engaging national champions early, engaging in two-way dialogue designed to maximise practical usefulness of results, and ownership and understanding of results in national champions. The co-applicants on this proposal are the first set of national champions, who are opinion leaders in their fields and sit on the national decision-making bodies. Co-applicants will work with colleagues through national groups and beyond adapting and optimising the communication strategy to communicate to their peers, and receive feedback at every stage of the research. The groups we will involve include the English Breast Clinical Advisory Group (Wallis/Given-Wilson/Taylor-Phillips/Kearins/Pinder members), the Clinical and Professional Groups for Breast Radiology (Given-Wilson/Wilkinson member), Radiography, administration, Breast Care Nursing and Surgery, the National Co-ordinating Committee for Breast Pathology (Pinder Chair) and the Breast Screening Advisory Committee (Wilkinson Chair), NHS England Breast Screening Programme Board (Kearins member) as well as to the internal PHE Breast Screening Joint Action Meeting (Kearins member).

We will hold a workshop at the beginning and end of the project for national champions and key stakeholders. We will approach stakeholders through the existing national programme advisory network, at national conferences and regional breast screening professional network meetings, using the study teams extended networks, and via direct communication to the director of breast screening at each English Centre. We will engage a wide range of stakeholders using these methods, and expand our team of national champions.

The benefits and harms of breast cancer screening is a controversial area with substantial debate around research methods. We will contact prominent methodologists and researchers with a range of views about breast screening at a very early stage. We will seek feedback on our protocol, and engage with them in order to maximise the probability of acceptance within the scientific communities involved.

3. Communication to the public

We recognise that communication to the public will be challenging as this is a complex, controversial and emotive topic. Previous studies have used citizen's juries to engage members of the public for a prolonged period to give them time and resources to understand the complex benefits and harms of breast screening.(29) This has been successful, but has demonstrated that success in this context requires large amounts of people's time and significant financial resources,(29) and is beyond the scope of this project. In this context our objectives for communication to the public are firstly to clearly communicate how and why we are using women's data, and secondly to minimise the chances of misunderstanding of our results.

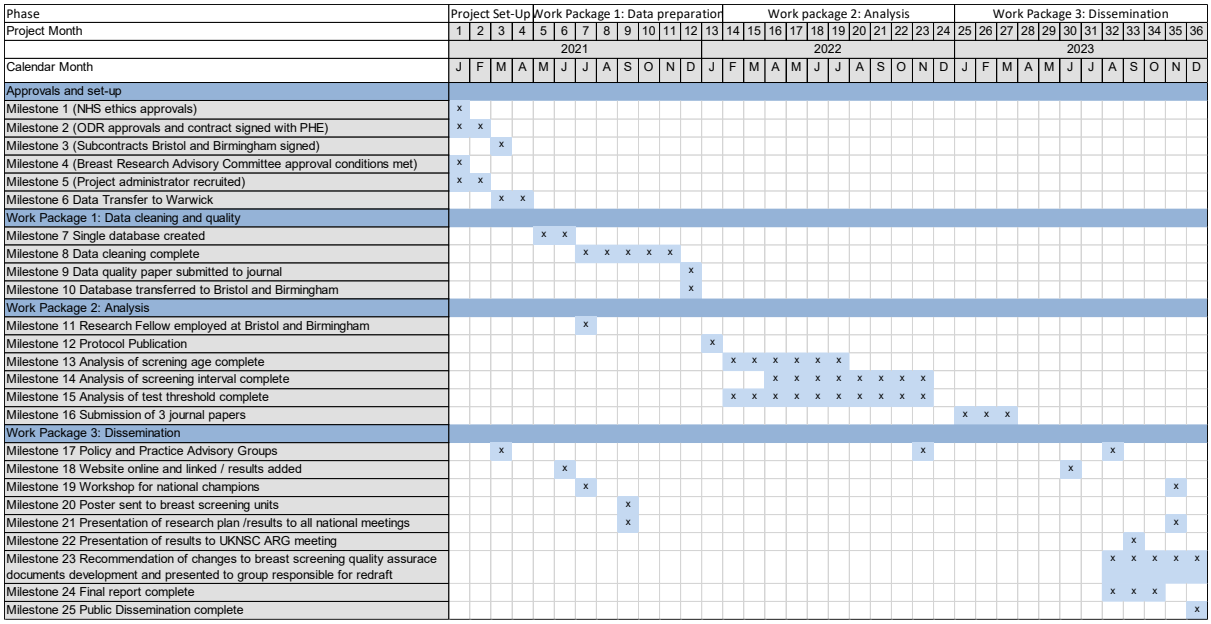
Communicating how and why we are using women's data. The PPI team will lead the design of a poster to send to all breast screening units to advise how data is used, direct them to the study website, and include details of how to opt out of future research, all from the patient's perspective. The co-applicants (Kearins/Wallis/Wilkinson/Given-Wilson) will assist with accuracy of content. The ATHENA-M website will be co-produced between ProfessorTaylor-Phillips, the PPI team, and the co-applicants. It will include sections targeted at clinicians and the public. Example content will be stories explaining why the research is important from different perspectives such as the PPI team and clinicians, how

people's data has been used and the ethical aspects, and findings as they emerge. This same content will be linked to the Independent Cancer Patient Voices website.

The possibility of misunderstanding our results will depend on the results themselves. For example, if our results suggest that the previous expansion to age 70 significantly increased the harms of breast cancer screening with few benefits there is potential for press attention and sensationalism. A potential misunderstanding would be that this means that women should not attend screening at all. To reduce these risks our strategy will emphasise maintaining control of the messaging, focusing on core messages which have been tested to minimise the chances of misunderstanding. ProfessorTaylor-Phillips will lead the science communication, with our PPI team and radiologist co-applicants communicating what this means for patients and the NHS. We will work closely with University of Warwick and PHE communications team during the project, to synchronise messages, as we have for previous projects. The University of Warwick also has an established relationship with the Science Media centre (an independent organisation aiming to make coverage of science more balanced), and they can assist with expert reaction to controversial stories. The core messaging will be centred on improving breast screening accuracy using 30 years of NHS experience. We will work extensively with Warwick University PPI volunteers (a large and diverse group of members of the public) and with members of Independent Cancer Patients' Voice (ICPV) to reduce the possibility of misunderstanding, or unforeseen consequences of our communication with people who are not familiar with the project. In this process we will seek a broad range of perspectives from groups of different ages, sex, ethnicity, and education. We will take particular care when communicating results concerning inequalities.

Study Management

The project timings are shown in the gantt chart below:



There will be monthly project management meetings within each University individually. All three universities will meet at least every 3 months via online meeting space. All co-applicants and team members will meet in person annually.

ProfessorTaylor-Phillips will lead the project and provide overall management. She will be mentored in this by Professor Aileen Clarke and Professor James Mason, both of whom have extensive experience in delivering large research projects across several universities. ProfessorTaylor-Phillips is already mentored by Professor Janet Dunn, Professor of Cancer Clinical Trials at Warwick Clinical Trials Unit.

There will be a postdoctoral research fellow at each of Warwick, Birmingham and Bristol Universities to carry out the work. Work package 1 will be undertaken primarily at Warwick, with support from Public Health England. Work package 2 analyses of age extension and screening interval will be undertaken at Bristol. Work package 2 test threshold will be undertaken primarily at Birmingham. However analysis of the overdiagnosis outcome will be led by ProfessorTaylor-Phillips and Karoline Freeman from Warwick. Professor Taylor-Phillips already works part time at Birmingham and Karoline Freeman holds an honorary contract there so it is an established collaboration. Work Package 3 will be led by Warwick with heavy involvement from Public Health England and all co-applicants.

There will also be a member of administrative staff employed at Warwick University responsible for coordinating the work between the three universities and Public Health England, and assisting with the administrative tasks involved in achieving all of the necessary approvals and documentation required in routine data projects.

There will be a policy and practice advisory group who meet in person every year. They will guide the research to maximise policy and practice relevance, and guide work package 3. We will also report progress updates to this group.

Appendix 1 – Letter of Support [redacted]

Appendix 2: Data requested and reason

Table 1: National Breast Screening Research Dataset (NBSRD) Patient Demographics data to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria.

To include one row per woman included, with pseudonymised identifiers to link between tables, with ethnicity, month and year and cause of death, month and year of birth, participation in relevant research trials, and issues with data quality of NHS number.

Table 2: National Breast Screening Research Dataset (NBSRD) breast screening episode data for routine population screening episodes to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria

To include one row per screening episode, with pseudonymised identifiers to link between tables, screening date, pseudonymised identifiers for the radiologists examining the mammograms, their decisions, whether the woman was recalled for further tests, whether cancer was detected.

Table 3: National Breast Screening Research Dataset (NBSRD) mammographic features to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria.

Details of mammographic features associated with detected cancers. Data will be provided for screen detected breast tumours only. Features to include side of the body, mammographic characteristics such as mass or calcifications.

Table 4: National Breast Screening Research Dataset (NBSRD) cancer tumour data to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria.

Tumours will be limited to C50 (breast cancer) or D05 (Breast DCIS or LCIS) records, or the pre-1995 equivalents in of ‘174’ and ‘2330’ respectively, only, inclusive of both screen detected and symptomatically detected breast tumours. Includes ICD classification, morphology, behaviour, grade, size, number of involved nodes, oestrogen, progesterone and HER2 status, Nottingham Prognostic Index, TNM stage, and whether screen detected.

Table 5: National Breast Screening Research Dataset (NBSRD) cancer treatment data to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria.

Treatment detailed will be restricted to records for the treatment of C50 or D05, or the pre-1995 equivalents in of ‘174’ and ‘2330’ respectively, (inclusive of screen detected and symptomatically detected breast tumours). Includes all events those occurring within the first 365 days from DATEOFDIAGNOSISBEST. All other event dates will be excluded. Treatment includes chemotherapy, radiotherapy and surgery.

Table 6: National Breast Screening Research Dataset (NBSRD) derived 'First Event' data to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria.

Events will be restricted to first procedure NCRAS is aware of for a C50 or D05, or the pre-1995 equivalents in of '174' and '2330' respectively, within the first 365 days from DATEOFDIAGNOSISBEST. All other event dates will be excluded. Datasources include registration data supplemented with SURGERYETC and OTHERPROCEDURES; HES; SACT; RTDS.

All dates and events for this table to be taken from multiple sources as described in the 'Custom fields' tab of the data dictionary. Includes breast surgery (breast conserving, mastectomy) underarm surgery (axillary clearance, sentinel lymph node biopsy), hormone therapy, radiotherapy, chemotherapy.

Table 7: National Breast Screening Research Dataset (NBSRD) breast screening episode data for non-routine screening episodes (high risk, GP/self-referral, non-routine recall) to be provided to the applicant in accordance with the cohort inclusion and exclusion criteria

To include date of any appointments for mammography which were not defined as screening appointments, such as GP referral.

Table 8: IMD Score – to be based upon last known address in both NCRAS and NBSS.

Index of multiple deprivation for woman's last known address only.

Table 9: Cause of date from the civil registration mortality file.

Month and year of death and cause for main study outcome. Cause primarily to separate breast cancer death from other causes, with sensitivity analyses defining this using combinations of underlying and secondary cause. ICD10 code C50 in underlying cause is main indicator for breast cancer detch. However ICD10 code C76 multiple cancers may include some breast cancer detahs, to check this will use ICD10 secondary cause, and refer death certificate data death cause code 1a,b,c and 2 to clarify.

Table 10 Cause of death from cancer registry

Month year and cause of death from the cancer registry, to check data quality of data from civil registration mortality file (where quality may have been limited by only linking on NHS number).

Table 11 Reference table with no data to extract

Table 12

Should the analysis expand in mutual agreement between members of the study team from Public Health England and Warwick, the following variables may also be required. Tumour histories of other (non-breast) cancers in women in the included cohort (as a potential confounder to the analysis). HES data providing additional detail of the morbidity associated with overdiagnosis and overtreatment (eg cardiac toxicity) and Charleston co-morbidity index (to characterise the population affected by overdiagnosis). Endocrine treatment data (not currently available as dispensed from community pharmacy, but linkage into cancer registry underway). Description of genomic characteristics of the cancer, which alongside other characteristics such as grade and stage may be important for predicting the benefits and harms of detecting each type of cancer in the mechanism of action analysis.

Appendix 3 Inclusion criteria

Women need an episode record (ODR Table 2) with date of first offered appointment (DOFOA) and age at DOFOA recorded and at least 1 demographics record (ODR Table 1) in order to be selected for the ATHENA-M cohort. So the ATHENA-M proposed cohort we are currently working towards is as follows (all dates are given for main project extract, there may be an initial extract with earlier dates and a subsequent extract with updated dates):

Women invited to routine population breast cancer screening in England from screening programme inception in 1988 to December 31st 2016, who:

(i) Were aged 47 – 73 years at their routine screening invitation

(ii) Have at least one demographics record, required for linking

The inclusion criteria by table are as follows (again for the main project extract)

- *Table 1: NHS National Breast Screening Programme (Patient Demographics) NBSS data will be restricted to women meeting the inclusion criteria.*
- *Table 2: NHS National Breast Screening Programme (routine episode) data will be restricted to women in Table 1. Records with no invitation to screening will be excluded from Table 2. Table 2 will be restricted to records with date of first offered appointment either (i) null or (ii) from screening programme inception in 1988 to 31/03/2018.*
- *Table 3: NHS National Breast Screening Programme (assessment procedures - mammographic features) NBSS data will be restricted to women in Table 1 and episodes in Table 2.*
- *Table 4: Cancer Registry data will be restricted to C50x or D05x records, or the pre-1995 equivalents in of '174' and '2330' for all women identified in Table 1, irrespective of whether the tumour was screen detected, where DATEOFDIAGNOSISBEST is between 01/01/1988 and 31/12/2018.*
- *Table 5: Cancer Registry data will be restricted to events linked to the tumours identified in Table 4 and where the event occurs within the first 365 days from DATEOFDIAGNOSISBEST and the OPCS4_CODE is identified in Table 10: Reference table OPCS4_codes. All other event data will be excluded.*
- *Table 6: Cancer Registry Derived 'First Event' data restricted to first procedure of each type the Cancer Registry is aware of for a C50x or D05x or the pre-1995 equivalents in of '174' and '2330' respectively, within the first 365 days from DATEOFDIAGNOSISBEST in Table 4. All other event data will be excluded. Data sources include registration data supplemented with SURGERYETC and OTHERPROCEDURES; HES; SACT; RTDS. All dates and events for this table to be taken from multiple sources as described in the 'Custom fields' tab of the data dictionary.*
- *Table 7: NHS National Breast Screening Programme (non-routine episode) data from BS Select will be restricted to women in Table 1. Records with no invitation to screening will be excluded from Table 7. Records in Table 7 will be restricted to records with date of first offered appointment either (i) null or (ii) from screening programme inception in 1988 to 31/03/2018*
- *Table 8: IMD Score data will be based upon last known address in NBSS.*
- *Table 9: MBIS civil registration mortality data be provided for all women identified in Table 1 and recorded in MBIS as deceased, updated in 2021.*
- *Table 10: NCRAS death data to be provided for all women identified in Table 4 and recorded in NCRAS as deceased*